AFRL-HE-AZ-TR-2004-0090

United States Air Force Research Laboratory

THE EFFECTIVENESS OF A TRADITIONAL GRADESHEET FOR MEASURING AIR COMBAT TEAM PERFORMANCE IN SIMULATED DISTRIBUTED MISSION OPERATIONS

Michael Krusmark

L-3 Communications
6030 South Kent Street
Mesa AZ 85212-6061

Brian T. Schreiber

Lockheed Martin
6030 South Kent Street
Mesa AZ 85212-6061

Winston Bennett Jr.

Air Force Research Laboratory
Warfighter Readiness Research Division
6030 South Kent Street Mesa, Arizona 85212-6061

May 2004

Human Effectiveness Directorate
Warfighter Readiness Research Division
6030 South Kent Street
Mesa, AZ 85212-6061

NOTICES

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of STINFO exchange.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public.

This paper has been reviewed and is approved for publication.

WINSTON BENNETT JR.
Project Scientist

HERBERT H. BELL
Technical Advisor

CURTIS J. PAPKE, Colonel, USAF
Chief, Warfighter Readiness Research Division

Contract Number: F41624-97-D-5000
Contractor: L-3 Communications

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
21-06-2004	Final	March 2003 to March 2004

4. TITLE AND SUBTITLE
The Effectiveness of a Traditional Gradesheet for Measuring Air Combat Team Performance in Simulated Distributed Mission Operations

5a. CONTRACT NUMBER
F41624-97-D-5000

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Michael Krusmark
Brian T. Schreiber
*Winston Bennett Jr.

5d. PROJECT NUMBER
1123

5e. TASK NUMBER
AS

5f. WORK UNIT NUMBER
03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

L-3 Communications
6030 South Kent Street
Mesa, AZ 85212-6061

*Air Force Research Laboratory
Warfighter Readiness Research Laboratory
6030 South Kent Street
Mesa, AZ 85212-6061

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Readiness Research Division
6030 South Kent Street
Mesa AZ 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL; AFRL/HEA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-AZ-TR-2004-0090

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
Air Force Research Laboratory Technical Monitor: Dr. Winston Bennett Jr., DSN: 474-6297, COMM: (480) 988-6561 Ext. 297

14. ABSTRACT

The utility and effectiveness of simulation-based training of air combat skill has been supported by a long research history (Bell & Waag, 1998). However, despite this success, there has been much debate on the effectiveness of gradesheets that have been used as the primary measure of air combat skill. The purpose of the current research was to assess the effectiveness of a Gradesheet that has been used to measure air combat team performance in Distributed Mission Operations (DMO). The current research assumed that DMO training is effective at training air combat skill, and assessed the DMO Gradesheet accordingly. It was hypothesized that the DMO Gradesheet is an effective measure of air combat team performance, and is therefore sensitive to differences in team performance over time during training, among performance indicators, and across team experience levels. Between August 2000 and December 2001, thirty-two teams of F-16 pilots from various United States Air Force operational units participated in week of structured DMO training. Training consisted of teams flying four networked high-fidelity F-16 simulators in scenarios against multiple constructed threats. On all scenarios of the training syllabus, air combat Subject Matter Experts (SMEs) graded performance of the teams on 40 indicators that comprise the DMO Gradesheet. Analyses of the data found that aggregate mean graded team performance increased over a week of DMO training. However, for all indicators, mean graded performance increased linearly and at approximately the same rate, suggesting that the Gradesheet captured only a single component of air combat performance. Furthermore, team experience did not moderate change in graded performance over missions as expected, undermining the claim that the Gradesheet measured even general performance. Thus, results suggest that the DMO Gradesheet as used in the current study lacked the sensitivity, the validity, and the reliability desired in a measure of air combat performance. Future directions in the development of subjective and objective measures of air combat performance are discussed.

15. SUBJECT TERMS
Air-to-air combat; distributed mission operations; instructor ratings; performance measurement; training evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Mr. Don Beedle
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED	80	**19b. TELEPHONE NUMBER** *(include area code)* 480.988.6561 x-552 DSN 474-6552

Table of Contents

List of Figures

List of Tables

Introduction

The purpose of the current research is to assess the effectiveness of the DMO (Distributed Mission Operations) "Gradesheet," a subjective measurement instrument that has been used at the Air Force Research Lab (AFRL) in Mesa to measure team air-to-air combat skill. Research on training of air-to-air combat skill in a simulated environment has a long history. Results from this research tell an encouraging and compelling story about the utility and effectiveness of simulation based training in one-aircraft or multi-aircraft platforms (Bell & Waag, 1998). However, even as the consensus has been that simulation based training is effective; there has been much debate on the reliability, validity, and sensitivity of instruments that have been used to measure air combat knowledge and skill. That the Gradesheet, in one form or another has been included in this debate is the primary motivator for the current assessment of its effectiveness. In choosing a methodology to assess the effectiveness of the Gradesheet, the present research was constrained by limited time and resources, which led us to base our evaluation on an analysis of archival Gradesheet data collected at the AFRL in Mesa between August 2000 and December 2001. During this time period, F-16 pilots from various United States Air Force operational units participated in a week of DMO training. This consisted of pilots flying as teams in four networked F-16 simulators against constructed multi-threat scenarios. For each team, on each scenario, Subject Matter Experts (SMEs) observed performance of 4-ship pilot teams, and assigned grades on selected indicators of team performance using the Gradesheet.

The current evaluation of the Gradesheet assumes that DMO training is effective at training air-to-air combat skill. Therefore, we begin with a brief discussion of relevant research that supports this assumption. Included in this discussion are results of a questionnaire administered to air combat SMEs that surveyed their beliefs of how Mesa DMO affects performance of 4-ship teams as they progress through a week of DMO. Next we review challenges that researchers have had over the years measuring air combat skill, as well as results that have led us to question the sensitivity, reliability, and validity of the current Gradesheet. This is followed by a discussion of the method we used to evaluate the effectiveness of the Gradesheet, given constraints that we had. Results are then presented and interpreted. Finally, we conclude with a discussion of current research collaborations with the AFRL in Mesa that are resulting in new measurement instruments of team air combat skill.

Research on Within-Simulator Learning in Air-to-Air Combat

Much research has been conducted over the last 30 years documenting the utility and effectiveness of simulation based training. Review of this research shows that utility evaluations have been the most common research methodology used to evaluate training in simulated air combat environments (Bell & Waag, 1998). Numerous research endeavors on multiple aircraft platforms have found that most pilots rate simulator based training as valuable (Bell & Waag). This is especially true of recent evaluations that have benefited from simulated training environments that are increasingly representative of a real-world air combat environment (Crane, Robbins, & Bennett, 2000). Although utility evaluations can be, and have been, successful at gauging pilot opinion of air combat training in a simulated environment, they are far from

convincing as a methodology for assessing change in performance during a simulated air combat training curriculum. Within-simulator evaluations are a more convincing methodology for assessing training effectiveness (Bell & Waag). The typical design employed in within-simulator learning research has been pre-post assessments of pilot performance in simulated air combat scenarios. Results from this research show that air combat performance improves across benchmark scenarios flown at the beginning and end of a training curriculum (for reviews, see Bell & Waag, 1998; Brecke & Miller, 1991; or Kelly, 1988).

Since the late 1980s, the nature of training in a simulated environment has undergone major transformations. The primary change has resulted from advances in technology that have allowed for training events to occur on a much greater scale. No longer is training relegated to stand-alone simulators. The current training environment at the AFRL is comprised of four high-fidelity F-16 simulations and an advanced warning and control system (AWACS) station linked together in an environment that allows for multiple types of 4 v. X scenarios. Moreover, this platform has been successfully employed in training events that have simultaneously linked the 4-ship in Mesa to multiple simulated aircraft at different locations worldwide in a synthetic Red Flag Training Range environment (Crane, 1999).

Several research projects have documented positive effects of within-simulator training of air-to-air combat in a distributed 4-ship training environment. In research conducted at the AFRL, Crane, Robbins, and Bennett (2000) evaluated the effectiveness of a DMO curriculum that was designed to augment training of pilots who were participating in Flight Lead Upgrade (FLUG) Training at their home units. The simulated training environment consisted of four high-fidelity F-16 simulators that were networked for simulated air combat engagements against constructed multi-threat presentations. SMEs graded relative performance of six pilots participating in FLUG-DMO on benchmarks at the beginning and end of the week-long FLUG-DMO syllabus. Performance was graded by SMEs using a 3 point scale in which 0, +, and – indicated average, above average, and below average mission performance for an upgrading FLUG pilot. Results show that flight leads demonstrated improvement between pre to post-training assessments of rated mission performance (Crane et al.).

More recently, the effectiveness of DMO has been assessed using data generated from an objective measurement system that is being developed at the AFRL in Mesa (Schreiber, Watz, & Bennett, 2003). During 2002, nineteen teams participated in a week of DMO using the networked F-16 simulator environment. Teams flew 4 v. 8 point-defense benchmark scenarios at the beginning and end of the DMO syllabus. Performance was measured using the Performance Measurement Tracking System (PETS), which passively collects data from the synthetic DMO environment on multiple process and outcome measures of performance. Analyses of data colleted using PETS indicated dramatic improvement in both summary process and outcome measures of 4-ship air combat performance, as reported in percentage changed (Schreiber et al.).

Predictions from SMEs on the Effectiveness of DMO for Select Indicators of Performance

In the history of research on within-simulator learning of air combat skill, little evidence has been collected assessing how specific skills necessary for successful air-to-air combat change

over missions of a training syllabus. For the most part, process measures of performance have been composites that summarize overall performance on a given engagement. When performance data has been collected on specific processes, analyses have focused on performance aggregated from all indicators (Seaman, 1999). Thus, to get a better understanding of how different air combat skills change over missions of a DMO syllabus, we conducted a study that asked F-16 SMEs to use their expertise to generate predictions about how Mesa DMO affects performance of 4-ship teams as they progress through a week of DMO. Moreover, because air combat flight experience has been found to consistently predict outcomes of air combat engagements (Waag & Houck, 1995), SMEs were asked if the impact of DMO on performance is different for teams that have more or less F-16 experience. Appendix A has the method and results of this study, and Appendix B has the full questionnaire used to elicit predictions from SMEs. The questionnaire asked SMEs to consider in turn each of 40 indicators of 4-ship team performance, and choose the pattern of performance that best represented their judgment of how quality of team performance progresses during a DMO training week for more and less experienced teams. To make their judgments, SMEs chose among eight patterns of performance that depicted graphically different performance patterns that might occur for 4-ship teams with more or less flight experience. The 40 indicators came from the DMO Gradesheet, and were defined in the questionnaire (Appendix B).

Predictions generated by SMEs through the questionnaire suggest that air combat team performance varies by indicator. First, performance was expected to improve with DMO training for many indicators, but not all indicators. For 37 of 40 indicators, if we look at performance collapsed across team experience levels, the majority of SMEs predicted that performance improves as teams proceed through a week of DMO training. However, for two indicators, Visual Lookout and Clear Avenue of Fire, average performance was not expected to change with training. Second, SMEs expected that team experience would moderate performance gain across missions for some indicators, but not for others. For 24 indicators of performance, the majority of SMEs predicted that team F-16 experience moderates change in performance over missions. For nine other indicators, the majority of SMEs predicted that team experience does not differentially affect change in performance. Third, for those indicators that SMEs predicted team experience as a moderator of performance, less experienced teams were expected to benefit more from training than more experienced teams. That is, for 17 indicators, the majority of SMEs predicted performance of less experienced teams to improve at a faster rate than more experienced teams. In summary, predictions made by SMEs suggest that DMO has greater training benefits for some skills relative to others.

Measures of Air combat Performance

Although research suggests that simulation based training is effective at improving air combat skill, and that the relative effectiveness of DMO training differs by performance indicator, success at documenting training effects has been limited in part by researchers' success at measuring air combat skill. Over the past 35 years, many theoretical and empirical research endeavors have focused on measurement of air combat performance (Becke & Miller, 1991). Researchers have consistently reported success at reliably measuring air combat outcomes, which have included kill ratios, exchange ratios, and the like, which are relatively easy to observe reliably. However, when research has focused on process measures of performance, less

success has been documented. Typically process has been assessed by having SMEs provide summary ratings on numerous process indicators of engagement performance, the sum of which are often combined into a single composite measure for data analysis (Seaman, 1999). The resulting composite measures have provided little understanding into what specific processes are leading to successful outcomes over multiple training scenarios (Brecke & Miller, 1991), and have been questioned for their validity and reliability at measuring distinct processes required in air combat. In several lines of research, process measures of air combat performance were collected from simulators by computers in real time, which presumably allowed for a more reliable assessment of process. From this research, summary measures of both positional advantage and aircraft state were found to be related to engagement outcomes (Waag, Raspotnik, & Leeds, 1992), and were shown to improve between pre and post-training assessments (McGuinness, Bouwman, & Puig, 1982). Notwithstanding all the effort that has been invested in developing measures of air combat skill, performance measurement remains an active area of investigation.

Situation Awareness Rating Scale. The Gradesheet used in the current study evolved from measures used in a research program on situation awareness initiated in 1991 at the AFRL in Mesa (Waag & Houck, 1995; Waag, Houck, Greschke, & Raspotnik, 1995). The goal of this research was to define situation awareness, generate a measure of it, and then assess the relation between situation awareness and air combat performance (Waag & Houck, 1995). To this end, Houck, Whitaker, and Kendall (1993) conducted a cognitive task analysis of experienced fighter pilots to define situation awareness. Results of the task analysis led to development of the Situation Awareness Rating Scale, as described by Waag and Houck, (1995). The Situation Awareness Rating Scale consisted of seven general and 24 specific behavioral indicators that were judged to be essential components of successful situation awareness during air-to-air combat. From these 31 behavioral indicators that comprise the Situation Awareness Rating Scale, a composite measure was computed, which was characterized as a global measure of situation awareness. All results reported by the authors focused on this composite measure of situation awareness. That is, results for specific behavioral indicators were not reported. Three variants of the Situation Awareness Rating Scale were developed, one each for self, peer, and instructor ratings of situation awareness. For all three variants, pilot experience significantly predicted situation awareness (Waag & Houck, 1994). Waag, Houck, Greschke, and Raspotnik (1995) used the Situation Awareness Rating Scale to assess pilot performance during a week-long training syllabus in a simulated two-ship F-15 environment. They found that the average composite situation awareness score of pilots weighted for mission complexity improved across missions of the training syllabus. Although no inferential statistics were provided, pilots participating in the research confirmed the descriptive results by reporting that their situation awareness improved due to training.

Current Gradesheet. The current Gradesheet is an expanded version of the Situation Awareness Rating Scale that was reworked by researchers and experienced fighter pilots to include 40 process and outcome indicators of air combat performance. Indicators from the research by Waag and Houck (1994) are compared to indicators in the present research in Table 1. As can be seen, the Situation Awareness Rating Scale and the DMO Gradesheet have many of the same indicators.

Table 1. **Comparison of air combat performance indicators that are included in the current DMO Gradesheet and the Situation Awareness Rating Scale.**

Current DMO Gradesheet	Situation Awareness Rating Scale (Waag & Houch, 1995)
Radar Mechanics:	Radar
El Strobe Control	
Range Control	
Azimuth Control	
Utilizing Correct Mode	
Gameplan:	
Tactics	Developing Plan; Tactical Knowledge
Execution	Executing Plan
Adjusting Plan On-the-Fly	Adjusting Plan On-the-Fly
Tactical Intercepts:	
Formation	
Detection – Commit	Maintain Track of Bogeys/Friendlies
Targeting	Targeting Decisions
Sorting	Radar Sorting; Threat Prioritization
BVR Launch and Leave	Fire-point Decisions
BVR Launch and React	
Intercept Geometry	Analyzing Engagement Geometry
Low Altitude Intercepts	
AAMD:	Overall Weapons System Proficiency
RMD	
IRCM	
Chaff – Flares	Defensive Reaction (chaff, flares, maneuvering)
Communications:	
3-1 Communication	Quality (brevity, accuracy, timeliness)
Radio Discipline	Ability to Effectively Use Information
GCI Interface	Ability to use Airborne Warning and Control System (AWACS)/GCI
Additional Indicators	
Engagement Decision	Assessing Offensiveness/Defensiveness; Threat Evaluation
Spike Awareness	Interpreting Radar Warning Receiver (RWR)
E F & N Pole	
Egress – Separation	
Contracts	
ROE Adherence	
ID Adherence	
Post Merge Maneuvering	
Mutual Support	Mutual Support
Visual Lookout	Lookout (VSD, RWR, visual)
Weapons Employment	Weapons Employment
Clear Avenue of Fire	
Fuel Management	
Flight Discipline	Discipline
Situation Awareness	Time-sharing Ability; Spatial Ability; Integrating Overall Information
Judgment	Reasoning Ability
Flight Leadership – Conduct	Flight Management; Decisiveness
Briefed Objectives Fulfilled	
Overall Engagement Grade	
	Interpreting Vertical Situation Display
	Tactical Electronic Warfare System

Legacy Gradesheets. Not only is the current Gradesheet similar to the Situation Awareness Rating Scale, but it is also similar to legacy gradesheets that have been used for years to measure air combat performance at operational units throughout the Air Force (Seaman, 1999). In fact, the rating scale used in the current Gradesheet can be traced back to the Tactical Air Command regulation for Training Records and Performance Evaluations in Formal Flying Training Programs (TACR 50-31, 1983; Waag, Pierce, & Fessler, 1987). The regulation calls for performance to be measured using the following criteria:

N/A = Not applicable to this engagement.
D = Performance was unsafe.
0 = Performance indicates lack of ability or knowledge.
1 = Performance is safe, but indicates limited proficiency. Makes errors of omission or commission.
2 = Performance is essentially correct. Recognizes and corrects errors.
3 = Performance is correct, efficient, skillful and without hesitation.
4 = Performance reflects an unusually high degree of ability.

Thus, the current evaluation of the Gradesheet is especially appropriate given the similarities between it and legacy gradesheets that have been used to measure air combat performance at numerous operation units around the Air Force.

Research using the Gradesheet to Evaluate F-16 4-ship Team Performance

Has the current Gradesheet been effective at measuring 4-ship team performance? Bennett, Schreiber, and Andrews (2002) used the Gradesheet to evaluate performance of eleven teams of pilots who completed a week of DMO training in the 4-ship simulated F-16 environment at the AFRL in Mesa. Teams were graded by F-16 SMEs on several problem solving competencies using the Gradesheet. A descriptive analysis of performance data among indicators revealed little to no improvement in graded 4-ship team performance over missions of the week-long syllabus. One explanation for the slight increase in mean graded performance over missions may be increased scenario complexity over missions, which was not accounted for (Bennett et al.). At the time, a valid measure of scenario complexity was not available, so rather than use an existing measure of questionable validity, scenario complexity was purposefully not considered in the analysis. Nonetheless, the results do raise questions regarding the sensitivity of the Gradesheet for measuring competencies required during air combat, especially given the extremely positive reports received from pilots participating in DMO training.

The Present Research

The goal of the current research is to evaluate the effectiveness of the Gradesheet at measuring various aspects of team air combat performance. Waag, Pierce, and Fessler (1987) posited five requirements for evaluating air combat performance measures:

1. Definition requirement – criterions exist that define different levels of performance

2. Validity requirement – the measure is getting at what it is intended to measure, which is assessed by content validity, predictive validity, and face validity
3. Reliability requirement – the measure can be used reliably, as assessed through inter- and intra-rater reliability
4. Sensitivity requirement – is the measure sensitive to variation in performance at the desired grain size
5. Practicality requirement – the measure is easy to use and cost effective

The current research evaluates the Gradesheet with respect to these five requirements. The evaluation will focus primarily on validity, reliability, and sensitivity requirements. Because the definition and practicality requirements are qualitative in nature, they will be addressed in the discussion. Validity, reliability, and sensitivity requirements are inherently linked. If a measure does not reliably measure performance at the desired grain size, then it cannot be a valid measure of performance. If the Gradesheet is a valid measure of air combat performance, then it will predict changes in performance (a) across time, (b) among performance indicators, and (c) across experience levels:

Hypotheses

1. *The Gradesheet is sensitive to changes in performance resulting from time in DMO.* Based both on the long history of research on within-simulator training, and predictions that were generated by Mesa SMEs, the current evaluation assumes that air combat performance increases with time in DMO. If performance does improve with time in DMO, then the Gradesheet should be sensitive to these changes.

2. *The Gradesheet is sensitive to performance differences across team experience levels.* The Situation Awareness Rating Scale, from which the Gradesheet is derived, was found to be significantly related with pilot experience, wherein more air combat experience is related to better situation awareness. Thus, it follows that the current Gradesheet should capture effects of team experience on performance.

3. *The Gradesheet is sensitive to performance differences among indicators.* Indicators included in the Gradesheet were identified by Houck, Whitaker, and Kendall (1993) as distinct operational tasks required for air combat. Because the DMO syllabus is structured to progressively increase in complexity over missions, with different missions stressing different operational demands, we expect performance over missions to vary by indicator. This expectation is supported by predictions made by Mesa SMEs suggesting that performance over missions varies by indicator.

Method

Participants

One hundred forty-eight F-16 pilots formed 32 teams that participated in Mesa DMO between August 2000 and December 2001. Participants typically arrived at the Mesa research site in pre-existing teams from various Air Force operational units around the United States.

Most participant teams were comprised of five pilots (M = 5.28, SD = 0.70), who rotated in and out of the 4-ship Simulated Training Environment between missions in order to support training goals of individual pilots and teams.

At the Mesa DMO test bed, participant teams are scheduled to fly a training syllabus that consists of two missions a day, one in the morning and afternoon, for four and one-half days. Teams that flew at least seven missions were included in the present study. Between August 2000 and December 2001, 32 of 57 teams that participated in DMO flew seven or more missions. Although all participant teams flew a minimum of seven missions, groups still varied in the number of missions that they flew, for a variety of reasons. Thirteen teams had missing data from at least one mission: Five groups had no graded 4-ship performance on Mission 3 because SMEs were grading 2-ship teams. Four groups had missing data from one of Missions 4 through 7, presumably because SMEs were not available for grading. Lastly, two groups were missing data on Missions 8 and 9, and four more groups were missing data from Mission 9. These data were missing because participants in these groups had scheduling conflicts such that they were unavailable for a full week of training.

Repeated Participation. Thirty-one of the 148 pilots in our sample participated in Mesa DMO more than once. Seventeen pilots flew with two different participant teams between August 2000 and December 2001, nine of whom participated in two groups during the same calendar week to augment groups that needed an additional pilot for one, two, or three missions. Three more pilots flew with three different participant teams during the same time period. Of the 20 pilots who flew in multiple participant teams between August 2000 and December 2001, four had DMO experience prior to August 2000. In addition, eleven other pilots in our sample also participated in DMO before August 2000.

F-16 Experience. Participants varied in their F-16 experience when they arrived for Mesa DMO. Two measures assessed the relative experience of the F-16 pilots. First, participants reported the total number of F-16 hours they had coming into Mesa DMO, which ranged from 85 to 3100 hours (M = 984.65, SD = 691.06). The second measure of participants F-16 experience was their reported current qualification level: 1 = Wingman, 2 = 2-Ship Lead, 3 = 4-Ship Lead, 4 = Mission Commander, 5 = Instructor Pilot (M = 3.64, SD = 1.65).

Distributed Mission Operations

Between August 2000 and December 2001, DMO training research was structured using a building-block approach, in which teams flew a series of scenarios that progressively increased in complexity over the course of training week, such that teams had to have mastered skills earlier in the week in order to successfully complete scenarios later in the week (Crane, 1999). The training syllabi consisted of nine missions. Before and after each mission, teams participated in a briefing session that lasted approximately one hour, and a debriefing session that lasted approximately one and a half hours. The mission itself lasted about one hour, during which teams flew between three and eight different scenarios. More than 200 different scenarios were available to populate a DMO syllabus, with content of scenarios ranging from 2 and 4-ship Visual Identification, Sweep, Surface-Attack Tactics, Offensive Counter-Air, and Defensive Counter-Air. The number of constructed threats presented in these scenarios ranged from one to

twenty-one. Because the 4-ship team is the primary unit of analysis in the present research, data collected from 2 v. X scenarios are not included in data analyses presented here. Furthermore, during the first two and sometimes three missions of a DMO syllabus, teams spent much of their time flying two-ship scenarios that are intended to familiarizing pilots with the simulated F-16 4-ship environment, and provide low-level training. Data collected during familiarization missions are not included in the present research. Thus, although we included in the current research teams that flew seven or more missions, the first two missions from each team were excluded from analysis because these were primarily used for familiarization.

Scenario Complexity. During the time-period (2000-2001) that the current data was collected, each team did not get the same set of scenarios. Moreover, teams did not fly specific benchmark scenarios at predetermined intervals during the training curriculum. That scenarios varied in complexity across each participant's training syllabus creates methodological challenges for assessing the relation between time in DMO and performance. In past research, scenarios have been weighted by their complexity in an attempt to address these challenges (Waag, Houck, Greschke, & Raspotnik, 1995). In the present research, two candidate measures of scenario complexity were considered as options for weighting scenarios by complexity. First, F-16 SMEs rated each scenario's level of difficulty on a scale of 1 to 5, basing their ratings in part on the number of threats included in each scenario. However, on some engagements scenarios were augmented with additional threats, or even additional scenarios. When multiple scenarios were used for a single engagement, as was often the case, no structured measure of that Engagement's Level of Difficulty was defined. Thus, these difficulty ratings could not be used to weight scenarios. Second, for each engagement SME graders tallied the total number of threats presented. Because threat frequency from multiple scenarios on a single engagement can be easily computed, this may be the better of the two complexity measures available in the present research. However, the total threat frequency was not reliably measured in the present research, as evidenced by cases in the data-set where more kills than threats were observed. Apparently threats were often recorded incorrectly or not at all, especially when scenarios were augmented with additional threats. Because of the questionable reliability and validity of both of these candidate scenario complexity measures, the present research makes no adjustments for complexity of scenario across missions. It could be argued that this undermines the power of subsequent data analyses that were conducted; however, reports from SMEs suggest that their ratings were in part based on scenario complexity. That is, a grade of 2 for a 4 v. 4 scenario would not be the same as a grade of 2 for a 4 v. 8 scenario.

DMO Performance Assessment

Gradesheet. As previously described in the introduction, the Gradesheet was designed to measure F-16 4-ship team performance in simulated combat situations. The Gradesheet lists 40 indicators of air combat team performance. For each indicator, F-16 SMEs graded performance on a scale of 0 to 4, with additional options of dangerous (D) and not applicable (N/A). The complete Gradesheet is included in Appendix C. In practice, grades of N/A and D were not observed. Furthermore, for all performance indicators, a grade of 0 was made on less than 1% of all engagements, and a grade of 4 was made on less than 2% of all engagements. Overall, approximately 97% of grades were 1, 2, or 3.

Graders. Seven F-16 SMEs graded performance of the 32 participant teams. For each participant team, one SME graded performance on each engagement of each mission. For two missions of one participant team, and one mission of another, the primary SME grader was unavailable, and a second SME filled in. SMEs were active duty or reserve Air Force pilots with extensive experience in a variety of air combat aircraft. All were mission qualified in Air Force operational aircraft, and all were instructor pilots who graduated from the USAF Weapons Instructor Course. SMEs had on average more than 2,000 hours of F-16 flight time.

Grader Resources. Graders had a number of resources to aid their grading task. Graders attended all briefings, missions, and debriefings. During the mission, graders sat at an instruction operator station (IOS) where they could (a) get a gods-eye view of all entities, (b) watch a spliced signal of all the avionics in each F-16 cockpit, (c) see videos capturing views out the front window of each cockpit, and (d) hear all communication. During debrief, these same displays were again available for playback of the mission. Graders relied on observable behaviors from these resources to infer skill on each of the indicators listed on the gradesheet.

Inter-Grader Reliability. Four participant teams from November and December of 2000 had two SMEs concurrently grading their performance. These data allowed us to assess inter-grader reliability on four of the seven SMEs that graded performance in the current research. A total of ninety-four scenarios from twenty-three missions were simultaneously graded by two SMEs. To estimate inter-grader reliability on graded performance, intra-class correlation coefficients were computed from grades of the two SMEs on each indicator of graded performance, which are presented in Table 2. Inter-grader reliability (α) of graded performance varied greatly across indicators. For most indicators, the reliability coefficient was small, suggesting that the estimated reliability of graded performance on these indicators is questionable. The average α across indicators was .42 (SD = .13).

Table 2. **Intraclass Correlation Coefficients (α) computed from two SMEs grading each indicator of 4-ship team performance on all engagements.**

Indicator	α
Radar Mechanics:	
El Strobe Control	0.50
Range Control	0.31
Azimuth Control	0.41
Utilizing Correct Mode	0.37
Gameplan:	
Tactics	0.52
Execution	0.43
Adjusting Plan On-the-Fly	0.45
Tactical Intercepts:	
Formation	0.34
Detection – Commit	0.39
Targeting	0.37
Sorting	0.49
BVR Launch and Leave	0.37
BVR Launch and React	0.32
Intercept Geometry	0.42
Low Altitude Intercepts	0.75
AAMD:	
RMD	0.48
IRCM	0.40
Chaff – Flares	0.39
Communications:	
3-1 Communication	0.25
Radio Discipline	0.31
GCI Interface	0.42
Additional Indicators	
Engagement Decision	0.32
Spike Awareness	0.54
E F & N Pole	0.44
Egress – Separation	0.46
Contracts	0.38
ROE Adherence	0.28
ID Adherence	0.54
Post Merge Maneuvering	0.70
Mutual Support	0.40
Visual Lookout	0.40
Weapons Employment	0.32
Clear Avenue of Fire	0.40
Fuel Management	0.11
Flight Discipline	0.39
Situation Awareness	0.54
Judgment	0.23
Flight Leadership – Conduct	0.44
Briefed Objectives Fulfilled	0.60
Overall Engagement Grade	0.71

Results

For each participant team, aggregate means were computed from graded performance on all scenarios within each mission. Aggregate mean performance on missions was subsequently used in all data analyses. The decision to aggregate data to the mission level was based on several factors. First and foremost, the briefing and debriefing sessions that teams participate in before and after a mission provide a natural break in context that defines mission as a conceptually meaningful unit of analysis. Second, preliminary research using the current measurement system suggested that this system is not sensitive to changes in performance within a mission. Finally, because we aggregated across scenarios within a mission, we have a more reliable or stable measure of performance at the mission level.

A multilevel modeling approach was used to analyze the data. Multilevel modeling analyzes change in a two stage process. The analysis begins with what is commonly called an individual growth model, in which separate analyses are conducted for each participant team. Conceptually, the individual growth model is similar to an ordinary least squares (OLS) regression model that estimates linear growth trajectories by regressing tactics on mission progression for each participant team (Kenny, Bolger, & Kashy, 2002). That is, separate regression analyses would be computed for each participant team on each indicator to estimate (a) intercepts, graded performance of each team at the onset of DMT, and (b) slopes, change in graded performance over missions. For example, consider Figure 1 which depicts linear growth trajectories for 32 teams on estimated graded tactics. Graded tactics represent SMEs' judgments of how well 4-ship teams handled specific threat presentations in light of their mission briefings. Computationally, multilevel modeling techniques estimate linear change trajectories using maximum likelihood estimations, as opposed to OLS estimates in regression. Following analysis of the individual growth model, we proceed to the between-person model, which explores the possibility that intercepts and slopes vary systematically among teams. Between-person models can be used to explore contextual factors that may distinguish trajectories of different teams. For example, Figure 1 distinguishes among participant teams that have more or less experience – dashed green lines for less experienced teams, and solid blue lines for more experienced teams. For simplicity of presentation, team F-16 experience is depicted in Figure 1 as a dichotomous variable; however, in the analyses that follow, F-16 experience is treated as a continuous variable. For a thorough introduction to multilevel modeling, see Bryk and Raudenbush (1992) or Kreft and de Leeuw (1998).

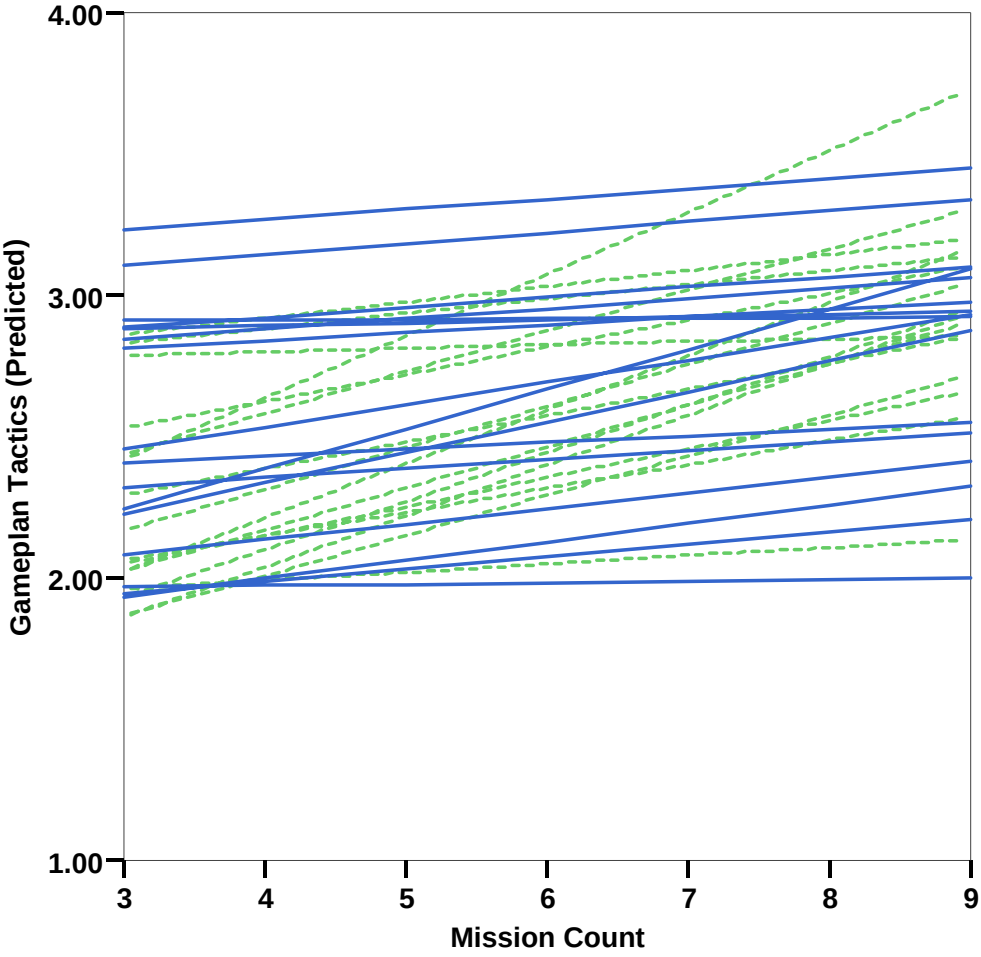


Figure 1. **Fitted linear change trajectories on graded tactics for thirty-two 4-ship teams of more and less experience. Blue solid lines represent teams above the median in weighted mean F-16 experience, while green dashed lines represent teams below the median.**

Change in Graded 4-ship Team Performance over Missions (Hypotheses 1)

For each indicator, estimates of the average graded performance at the onset of DMO, and the average linear change in graded performance over missions of DMO were computed using the Mixed Models option in SPSS. Table 3 presents these intercept and slope estimates for each indicator. First, consider the intercept estimates. Models were specified so that intercepts represent graded performance at Mission 3, which was the first non-familiarization mission for all teams. Estimates of average graded performance at Mission 3 ranged among indicators from 1.74 to 2.39. Thus, on all indicators performance begins with a grade of 2 when rounded to the nearest integer. Next, for each indicator, consider in Table 3 estimates of the mean slope of graded team performance trajectories over missions. Notice that on all indicators of performance, the mean slope was positive, indicating that mean performance of all teams increased across missions. On average, performance increased at rates ranging from .060 to .114 grades per mission, depending on the indicator. Thus, for none of the indicators does estimated performance reach an average grade of 3 at Mission 9. However, for all indicators, graded

performance increased at a significant rate. So, for example, estimated tactical performance increased by .081 grades per mission, or from 2.387 at the Mission 3 to 2.873 at Mission 9, which is a statistically significant improvement.

For two indicators of 4-ship team performance, Formation and E, F & N Pole, computations in SPSS failed to converge on a solution. Maximum likelihood computations follow an iterative process whereby successive estimations of parameter values are made until values are reached that change very little from iteration to iteration. Lack of convergence can be related to estimating coefficients that are close to zero, which is a plausible explanation in the present case.

The multilevel model used to estimate average intercepts and slopes for each indicator specified that participant teams can differ in their initial performance and in their rate of change over missions. Thus, this model provides estimates of the variance among teams on intercepts and slopes, as well as the covariance between intercepts and slopes. Table 4 presents variance and covariance parameter estimates computed when fitting this model to graded performance across missions for each indicator. Variance at Mission 3 tells us about the variability among participant teams in graded performance at Mission 3, or the onset of DMO training. For 34 of 40 indicators of performance, participant teams varied significantly on estimated graded performance at the beginning of the DMO syllabus, indicating that teams began with different baseline abilities for these indicators. Variance in mission progression presents us with information about variability in the slopes of the linear change trajectories from each team, and thus if there were differences among teams in their graded performance over missions. On only six indicators was there significant variability among 4-ship teams on the rate that graded performance improved over missions. These six indicators include three of four radar mechanics indicators: Range Control, Azimuth Control, and Utilizing Correct Mode, as well as SMEs grades of team skill at Detecting to Commit, Targeting, and Gameplan Tactics. Table 4 also presents estimates of the covariance between slopes and intercepts. Graded performance at Mission 3 was related to the rate of change of graded performance across missions on only one indicator of graded performance: how well teams selected the appropriate radar range. The final source of information about variability in team growth trajectories presented in Table 4 is residual variance, which is the variation within teams that is not explained by the model. For all indicators, significant variability remains within data from each team that is not explained by mission progression. Finally, returning to Figure 1, consider once again growth trajectories on graded Tactics that were estimated for each team using the unconditional linear growth model. Focusing on variability in performance among teams, it is easy to see that teams varied significantly on graded tactics both at the onset of training and over missions. In Appendix D, for each indicator of performance, separate charts show how performance trajectories varied among teams.

Table 3. **Mixed model parameter estimates of the effects of mission progression on performance for each indicator. Standard errors of estimates presented in parentheses.**

Indicator	Mission 3 (Intercept)	Mission Progression (Slope)
Radar Mechanics:		
El Strobe Control	2.1834*** (.0913)	.0757*** (.0170)
Range Control	2.2193*** (.1024)	.0865*** (.0184)
Azimuth Control	2.3258*** (.0947)	.0777*** (.0180)
Utilizing Correct Mode	2.2104*** (.1049)	.0952*** (.0182)
Gameplan:		
Tactics	2.3872*** (.0872)	.0813*** (.0173)
Execution	1.8713*** (.0828)	.0737** (.0221)
Adjusting Plan On-the-Fly	1.8584*** (.0783)	.0730** (.0208)
Tactical Intercepts:		
Formation	Failed to Converge	
Detection – Commit	2.2801*** (.0774)	.0712*** (.0181)
Targeting	2.1764*** (.0890)	.0640** (.0230)
Sorting	1.9781*** (.0789)	.0805*** (.0199)
BVR Launch and Leave	1.9852*** (.0754)	.0867*** (.0141)
BVR Launch and React	1.9406*** (.0768)	.0803*** (.0150)
Intercept Geometry	2.0407*** (.0730)	.0693** (.0180)
Low Altitude Intercepts	1.7427*** (.1074)	.1034*** (.0240)
AAMD:		
RMD	1.8499*** (.0971)	.0872*** (.0193)
IRCM	1.7386*** (.0754)	.0843*** (.0173)
Chaff – Flares	1.7990*** (.0859)	.1030*** (.0165)
Communications:		
3-1 Communication	1.8427*** (.1077)	.1137*** (.0164)
Radio Discipline	1.8582*** (.1130)	.0891*** (.0172)
GCI Interface	1.9518*** (.1053)	.1084*** (.0203)
Additional Indicators		
Engagement Decision	2.0938*** (.0739)	.0758*** (.0167)
Spike Awareness	2.0349*** (.0812)	.0931*** (.0151)
E F & N Pole	Failed to Converge	
Egress – Separation	1.9378*** (.0825)	.0745*** (.0178)
Contracts	2.0565*** (.0902)	.0905*** (.0172)
ROE Adherence	2.3206*** (.0749)	.0801*** (.0148)
ID Adherence	2.3166*** (.1027)	.0824*** (.0203)
Post Merge Maneuvering	1.8223*** (.0908)	.1011*** (.0177)
Mutual Support	1.8864*** (.0935)	.0844*** (.0195)
Visual Lookout	1.9602*** (.0817)	.0870*** (.0122)
Weapons Employment	2.0420*** (.0809)	.0813*** (.0163)
Clear Avenue of Fire	2.2570*** (.0924)	.0766*** (.0194)
Fuel Management	2.0785*** (.0803)	.0598** (.0202)
Flight Discipline	2.0967*** (.0797)	.0802*** (.0148)
Situation Awareness	1.8232*** (.0878)	.0696** (.0213)
Judgment	2.0128*** (.0783)	.0675** (.0191)
Flight Leadership – Conduct	2.1436*** (.0759)	.0691** (.0192)
Briefed Objectives Fulfilled	1.9440*** (.0779)	.0656** (.0214)
Overall Engagement Grade	1.7502*** (.0792)	.0893*** (.0182)

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 4. **Variance/Covariance parameters of random effects.**

Indicator	Residual Variance	Variance at Mission 3 (Intercept)	Variance in Mission Progression (Slope)	Covariance between Intercept and Slope
Radar Mechanics:				
El Strobe Control	.1134*** (.0133)	.2064** (.0684)	.0044 (.0023)	-.0109 (.0102)
Range Control	.0833*** (.0098)	.2905** (.0854)	.0072* (.0028)	-.0282* (.0131)
Azimuth Control	.0762*** (.0090)	.2455** (.0733)	.0070** (.0027)	-.0244 (.0118)
Utilizing Correct Mode	.1122*** (.0131)	.2917** (.0897)	.0057* (.0027)	-.0262 (.0133)
Gameplan:				
Tactics	.0818*** (.0096)	.1992** (.0623)	.0059* (.0024)	-.0192 (.0104)
Execution	.2052*** (.0241)	.1109* (.0564)	.0068 (.0041)	-.0170 (.0135)
Adjusting Plan On-the-Fly	.1934*** (.0227)	.0943 (.0522)	.0056 (.0036)	-.0141 (.0122)
Tactical Intercepts:				
Formation	Failed to Converge			
Detection – Commit	.0874*** (.0103)	.1440** (.0490)	.0066* (.0026)	-.0153 (.0096)
Targeting	.1412*** (.0166)	.1770** (.0636)	.0106* (.0043)	-.0233 (.0139)
Sorting	.1453*** (.0172)	.1211* (.0520)	.0063 (.0033)	-.0110 (.0108)
BVR Launch and Leave	.1403*** (.0167)	.1042* (.0476)	.0002 (.0017)	.0033 (.0070)
BVR Launch and React	.1435*** (.0177)	.1032* (.0500)	.0004 (.0019)	.0036 (.0076)
Intercept Geometry	.1312*** (.0156)	.1001* (.0452)	.0047 (.0027)	-.0099 (.0094)
Low Altitude Intercepts	.2320*** (.0288)	.2220* (.0925)	.0072 (.0050)	-.0221 (.0187)
AAMD:				
RMD	.2110*** (.0250)	.1804* (.0795)	.0025 (.0030)	-.0162 (.0137)
IRCM	.1847*** (.0240)	.0703 (.0546)	.0007 (.0030)	.0019 (.0106)
Chaff – Flares	.1896*** (.0224)	.1322* (.0629)	.0005 (.0023)	-.0043 (.0100)
Communications:				
3-1 Communication	.1234*** (.0145)	.3054** (.0956)	.0033 (.0023)	-.0140 (.0121)
Radio Discipline	.1479*** (.0175)	.3302** (.1062)	.0031 (.0025)	-.0116 (.0133)
GCI Interface	.1345*** (.0160)	.2820** (.0911)	.0073 (.0034)	-.0298 (.0153)
Additional Indicators				
Engagement Decision	.1470*** (.0173)	.0967* (.0447)	.0027 (.0023)	-.0071 (.0087)
Spike Awareness	.1543*** (.0180)	.1287* (.0544)	.0007 (.0019)	-.0055 (.0084)
E F & N Pole	Failed to Converge			
Egress – Separation	.1886*** (.0222)	.1180* (.0561)	.0021 (.0027)	-.0053 (.0100)
Contracts	.1178*** (.0140)	.1968** (.0664)	.0044 (.0025)	-.0126 (.0106)
ROE Adherence	.0878*** (.0104)	.1322** (.0458)	.0032 (.0018)	-.0066 (.0073)
ID Adherence	.1493*** (.0177)	.2566** (.0868)	.0067 (.0036)	-.0260 (.0153)
Post Merge Maneuvering	.1700*** (.0207)	.1686* (.0703)	.0026 (.0030)	-.0165 (.0129)
Mutual Support	.1923*** (.0225)	.1773* (.0716)	.0039 (.0031)	-.0129 (.0127)
Visual Lookout	.1127*** (.0132)	.1546** (.0544)	< .0000 (.0013)	.0015 (.0064)
Weapons Employment	.1565*** (.0182)	.1272* (.0535)	.0018 (.0021)	-.0053 (.0089)
Clear Avenue of Fire	.1503*** (.0178)	.1924** (.0701)	.0056 (.0033)	-.0219 (.0133)
Fuel Management	.1058*** (.0140)	.0993* (.0502)	.0060 (.0032)	-.0109 (.0103)
Flight Discipline	.0876*** (.0103)	.1562** (.0517)	.0032 (.0018)	-.0090 (.0076)
Situation Awareness	.2224*** (.0261)	.1282* (.0633)	.0051 (.0038)	-.0192 (.0141)
Judgment	.1953*** (.0229)	.0926 (.0513)	.0034 (.0030)	-.0106 (.0112)
Flight Leadership – Conduct	.1801*** (.0212)	.0887 (.0482)	.0042 (.0031)	-.0081 (.0107)
Briefed Objectives Fulfilled	.2136*** (.0254)	.0809 (.0524)	.0055 (.0039)	-.0101 (.0125)
Overall Engagement Grade	.2028*** (.0238)	.0947 (.0527)	.0022 (.0029)	-.0101 (.0111)

* $p < .05$; ** $p < .01$; *** $p < .001$

Effect of Team Experience on Change in Graded Performance across Missions (Hypothesis 2)

Having first estimated an unconditional linear change model, we then specified a model that included team F-16 experience. Results from the unconditional linear change model showed (a) significant variability on graded performance at Mission 3 for 34 indicators, and (b) significant variability on slopes of graded team performance trajectories for 6 indicators. We hypothesized that team F-16 experience explains variability among teams on intercepts and slopes for these indicators. In specifying a multilevel model that includes team F-16 experience as a predictor of graded performance, we defined team experience as a time-dependent covariate. Because most teams participating in Mesa DMO were comprised of more than 4 pilots that rotated in and out of the F-16 cockpit to accommodate the training needs of all pilots in the group, team F-16 experience is a time-dependent covariate.

Estimates of intercepts and slopes from a multilevel model that included team F-16 experience as a time-dependent predictor of graded performance are presented in Table 5. For ease of interpretation, team F-16 experience was centered at the grand mean of all teams on all missions, which resulted in estimates of mean intercepts and slopes that were the same for models with and without experience as a predictor. In Table 5, it is apparent that experience was a significant predictor of mean graded team performance at Mission 3 on six indicators of performance: Gameplan Tactics and Execution, Engagement Decision, Mutual Support, Judgment, and Flight Leadership. For these indicators, experience explains a significant proportion of the variability at Mission 3 that was found in the unconditional change model, wherein at Mission 3 more experienced teams have higher estimated graded performance than less experienced teams. Team experience was a significant predictor of the mean change in graded performance across missions on teams' decisions to egress and separate. So, teams with less experience demonstrated more improvement in graded egress and separation over missions than teams with more experience.

Table 6 presents variance and covariance parameter estimates for the linear change model that included team F-16 experience as a moderator of performance. Variance parameters estimates from this model are interpreted as the variability in intercepts and slopes remaining after controlling for team experience. As can be seen in Table 6, after controlling for team experience, significant variability among teams in graded performance is observed at the onset of DMO training for 27 indicators. Significant variability among teams on slopes of estimated graded performance trajectories was observed for six indicators, and these were the same performance indicators that showed significant variability among slopes estimated from the unconditional linear change model.

Table 5. **Mixed model parameter estimates of the effects of mission progression and team experience (F-16 hours) on performance for each indicator. Team experience is a time-dependent covariate.**

Indicator	Mission 3 (Intercept)	Mission Progression	Mission 3 x Experience	Mission Progression x Experience
Radar Mechanics:				
El Strobe Control	2.1738*** (.0880)	.0776*** (.0170)	.00035 (.00019)	-.000012 (.000044)
Range Control	2.2113*** (.0987)	.0881*** (.0181)	.00032 (.00019)	-.000042 (.000045)
Azimuth Control	2.3191*** (.0913)	.0792*** (.0176)	.00025 (.00018)	-.000039 (.000044)
Utilizing Correct Mode	2.2062*** (.1037)	.0961*** (.0181)	.00017 (.00021)	-.000027 (.000046)
Gameplan:				
Tactics	2.3779*** (.0823)	.0835*** (.0162)	.00038* (.00017)	-.000076 (.000041)
Execution	1.8617*** (.0762)	.0759** (.0212)	.00042* (.00020)	-.000065 (.000056)
Adjusting Plan On-the-Fly	1.8531*** (.0773)	.0742** (.0202)	.00025 (.00020)	-.000068 (.000053)
Tactical Intercepts:				
Formation	Failed to Converge			
Detection – Commit	2.2754*** (.0764)	.0721*** (.0183)	.00016 (.00017)	-.000002 (.000046)
Targeting	2.1722*** (.0870	.0650** (.0220)	.00019 (.00021)	-.000089 (.000055)
Sorting	1.9697*** (.0748)	.0823*** (.0192)	.00029 (.00018)	-.000069 (.000050)
BVR Launch and Leave	Failed to Converge			
BVR Launch and React	1.9379*** (.0767)	.0805*** (.0150)	.00014 (.00018)	-.000041 (.000042)
Intercept Geometry	2.0371*** (.0731)	.0699** (.0184)	.00014 (.00018)	-.000003 (.000048)
Low Altitude Intercepts	1.7312*** (.1058)	.1058*** (.0242)	.00028 (.00027)	-.000066 (.000065)
AAMD:				
RMD	1.8504*** (.0986)	.0870*** (.0196)	.00002 (.00024)	-.000014 (.000053)
IRCM	Failed to Converge			
Chaff – Flares	1.7953*** (.0862)	.1037*** (.0163)	.00026 (.00021)	-.000068 (.000045)
Communications:				
3-1 Communication	1.8395*** (.1076)	.1142*** (.0158)	.00019 (.00021)	-.000069 (.000042)
Radio Discipline	1.8582*** (.1142)	.0888*** (.0168)	.00010 (.00022)	-.000060 (.000045)
GCI Interface	1.9490*** (.1060)	.1091*** (.0207)	.00006 (.00023)	-.000034 (.000053)
Additional Indicators				
Engagement Decision	2.0840*** (.0697)	.0776*** (.0169)	.00036* (.00017)	-.000006 (.000045)
Spike Awareness	2.0311*** (.0810)	.0939*** (.0153)	.00016 (.00019)	-.000010 (.000042)
E F & N Pole	Failed to Converge			
Egress – Separation	1.9347*** (.0807)	.0747*** (.0168)	.00031 (.00020)	-.000105* (.000046)
Contracts	2.0582*** (.0918)	.0901*** (.0174)	.00003 (.00020)	-.000011 (.000045)
ROE Adherence	2.3154*** (.0727)	.0812*** (.0148)	.00018 (.00016)	-.000013 (.000039)
ID Adherence	2.3074*** (.0982)	.0844*** (.0199)	.00032 (.00022)	-.000031 (.000052)
Post Merge Maneuvering	1.8165*** (.0908)	.1024*** (.0178)	.00018 (.00021)	-.000017 (.000047)
Mutual Support	1.8755*** (.0900)	.0867*** (.0189)	.00044* (.00021)	-.000071 (.000051)
Visual Lookout	1.9536*** (.0818)	.0884*** (.0123)	.00025 (.00016)	-.000019 (.000034)
Weapons Employment	2.0417*** (.0819)	.0814*** (.0165)	-.00001 (.00019)	.000015 (.000045)
Clear Avenue of Fire	2.2489*** (.0911)	.0782*** (.0194)	.00034 (.00021)	-.000058 (.000051)
Fuel Management	2.0803*** (.0809)	.0583** (.0201)	-.00004 (.00021)	-.000063 (.000053)
Flight Discipline	2.0892*** (.0801)	.0817*** (.0151)	.00027 (.00017)	-.000012 (.000039)
Situation Awareness	1.8134*** (.0845)	.0718** (.0212)	.00035 (.00021)	-.000035 (.000056)
Judgment	2.0023*** (.0745)	.0698** (.0190)	.00040* (.00019)	-.000034 (.000051)
Flight Leadership – Conduct	2.1345*** (.0728)	.0709** (.0191)	.00038* (.00019)	-.000054 (.000051)
Briefed Objectives Fulfilled	1.9361*** (.0754)	.0674** (.0214)	.00029 (.00020)	-.000020 (.000056)
Overall Engagement Grade	1.7438*** (.0778)	.0908*** (.0182)	.00027 (.00020)	-.000023 (.000049)

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 6. **Variance/Covariance parameters of random effects controlling for team experience.**

Indicator	Residual Variance	Variance at Mission 3 (Intercepts)	Variance in Mission Progression (Slopes)	Covariance between Intercepts and Slopes
Radar Mechanics:				
El Strobe Control	.1126*** (.0132)	.1869** (.0639)	.0043 (.0023)	-.0087 (.0099)
Range Control	.0841*** (.0099)	.2757** (.0799)	.0067* (.0027)	-.0245* (.0123)
Azimuth Control	.0775*** (.0092)	.2244** (.0697)	.0064* (.0026)	-.0209 (.0114)
Utilizing Correct Mode	.1135*** (.0134)	.2827** (.0893)	.0055* (.0027)	-.0248 (.0134)
Gameplan:				
Tactics	.0831*** (.0098)	.1719** (.0562)	.0047* (.0022)	-.0135 (.0092)
Execution	.2080*** (.0246)	.0759 (.0486)	.0055 (.0040)	-.0098 (.0120)
Adjusting Plan On-the-Fly	.1954*** (.0230)	.0881 (.0514)	.0047 (.0034)	-.0117 (.0118)
Tactical Intercepts:				
Formation	Failed to Converge			
Detection – Commit	.0877** (.0104)	.1384** (.0484)	.0068* (.0028)	-.0148 (.0097)
Targeting	.1423*** (.0169)	.1647** (.0633)	.0093* (.0040)	-.0183 (.0132)
Sorting	.1481*** (.0176)	.0996* (.0482)	.0054 (.0031)	-.0063 (.0100)
BVR Launch and Leave	Failed to Converge			
BVR Launch and React	.1448*** (.0178)	.1010* (.0504)	.0003 (.0019)	.0040 (.0076)
Intercept Geometry	.1314*** (.0156)	.1003* (.0457)	.0051 (.0029)	-.0108 (.0097)
Low Altitude Intercepts	.2339*** (.0291)	.2068* (.0913)	.0073 (.0051)	-.0205 (.0186)
AAMD:				
RMD	.2120*** (.0252)	.1876* (.0823)	.0027 (.0031)	-.0173 (.0142)
IRCM	Failed to Converge			
Chaff – Flares	.1902*** (.0225)	.1329* (.0630)	.0002 (.0022)	-.0036 (.0098)
Communications:				
3-1 Communication	.1247*** (.0147)	.3038** (.0960)	.0027 (.0022)	-.0129 (.0118)
Radio Discipline	.1496*** (.0177)	.3373** (.1088)	.0026 (.0025)	-.0121 (.0133)
GCI Interface	.1333*** (.0159)	.2859** (.0925)	.0078 (.0036)	-.0305 (.0157)
Additional Indicators				
Engagement Decision	.1459*** (.0172)	.0775 (.0404)	.0029 (.0024)	-.0029 (.0024)
Spike Awareness	.1545*** (.0181)	.1269* (.0543)	.0009 (.0019)	-.0056 (.0086)
E F & N Pole	Failed to Converge			
Egress – Separation	.1899*** (.0224)	.1075* (.0544)	.0010 (.0024)	-.0022 (.0091)
Contracts	.1183*** (.0142)	.2052** (.0712)	.0045 (.0027)	-.0135 (.0114)
ROE Adherence	.0888*** (.0106)	.1211** (.0445)	.0031 (.0019)	-.0052 (.0072)
ID Adherence	.1510*** (.0181)	.2263** (.0815)	.0061 (.0036)	-.0210 (.0146)
Post Merge Maneuvering	.1700*** (.0208)	.1680* (.0710)	.0026 (.0031)	-.0160 (.0130)
Mutual Support	.1923*** (.0225)	.1562* (.0669)	.0032 (.0029)	-.0085 (.0117)
Visual Lookout	.1118*** (.0132)	.1548** (.0549)	.0001 (.0013)	.0020 (.0064)
Weapons Employment	.1569*** (.0183)	.1314* (.0557)	.0021 (.0022)	-.0061 (.0093)
Clear Avenue of Fire	.1501*** (.0178)	.1840** (.0685)	.0055 (.0033)	-.0207 (.0131)
Fuel Management	.1066*** (.0142)	.1007 (.0523)	.0057 (.0032)	-.0113 (.0107)
Flight Discipline	.0858*** (.0102)	.1585** (.0526)	.0035 (.0019)	-.0087 (.0077)
Situation Awareness	.2202*** (.0261)	.1110 (.0593)	.0050 (.0039)	-.0160 (.0135)
Judgment	.1929*** (.0230)	.0751 (.0467)	.0033 (.0033)	-.0081 (.0105)
Flight Leadership – Conduct	.1792*** (.0210)	.0743 (.0442)	.0040 (.0031)	-.0059 (.0100)
Briefed Objectives Fulfilled	.2127*** (.0254)	.0690 (.0485)	.0056 (.0041)	-.0081 (.0122)
Overall Engagement Grade	.2022*** (.0238)	.0877 (.0511)	.0022 (.0029)	-.0091 (.0109)

* $p < .05$; ** $p < .01$; *** $p < .001$

Quadratic Fits to the Data

Analysis was conducted to determine whether graded performance was better explained by curvilinear fits to the data. Results generated using the Mixed Procedure in SPSS were inconclusive, as computations consistently failed to converge on a solution for all indicators. In general, maximum likelihood estimations require a minimum of five participants per parameter. With 32 participant teams, and ten parameters in the quadratic model, it is not surprising that the model failed to converge.

So, to determine whether quadratic fits to the data better represent graded performance over missions, we relied on ordinary least squares estimations, following recommendations made by Kenny, Bolger, and Kashy (2002). First, for each participant team on each indicator, graded performance was regressed on mission count and mission count squared. This gave us regression estimates of both the linear and quadratic terms for all participant teams on all indicators, similar to the individual growth estimates in the maximum likelihood model described above. In between-person analyses, for each indicator t-tests were computed comparing the average linear and quadratic coefficients to zero. For all indicators, models with quadratic fits to the data added little information in explaining change in graded performance across missions. However, because quadratic effects were assessed through ordinary least squares estimations; they should be considered suggestive at best.

Consistency in Graded Performance among Indicators (Hypothesis 3)

Because the patterns of graded performance described above were found to be highly consistent across indicators, we wondered whether these data could be reduced to one or several components that could explain most of the variance in graded performance among indicators, a supposition motivated in part by research of Waag and Houck (1995). As previously described, their research involved the Situation Awareness Rating Scale, an earlier version of the current Gradesheet. Waag and Houck found that data from the 31 indicators that comprise the measure could be reduced to a single component that accounted for 92.5% of the variance.

In an attempt to get at whether a single factor explains most of the variance in graded performance on the forty indicators in the current research, we performed a Principal Components Analysis (PCA). In choosing PCA as a data reduction method, we were confronted with several challenges. It is generally accepted that one needs at least 300 cases to generate a stable factor solution, and that samples of 50 cases or less are very poor (Tabachnick & Fidell, 2001). A case is typically defined as a participant, or participant team in the present research. Because we had a sample of only 32 participant teams, we chose to consider each measurement occasion as a case. A measurement occasion was defined by the unit of analysis, which is the aggregate mean graded performance per mission. By defining a case as a measurement occasion for a given team, the analysis assumes that factors underlying the data are stable over measurement occasions. Because PCA makes no assumptions about independence of cases, we chose it as our exploratory factor analysis method.

To estimate the number of components that best describe graded 4-ship team performance, a PCA model was specified to extract components with eigenvalues greater than one. Four components had eigenvalues greater than one. However, one component explained most of the variance in graded performance over missions. Of the four components extracted, the first explained 62.47 percent of the variance in graded performance among indicators, while the additional three components explained 4.67, 4.39, and 3.01 percent of the variance respectively. Results of a scree test suggest that one factor be retained. Appendix E displays bivariate correlations among indicators on mean graded performance. Correlation coefficients ranged from .35 to .88, and were all significantly positive, $p < .001$. Based on the results of the scree test, a second PCA model was specified to estimate a one component solution. Appendix E also displays loadings of performance indicators on the principal component and communalities for the one component solution.

That the Gradesheet is capturing only one aspect of air combat skill is further supported by the internal consistency of graded performance among indicators. Cronbach's alpha was computed from graded performance at each measurement occasion for all 40 performance indicators. Cronbach's alpha was 98, which is very high. We would expect a much lower alpha had graded performance among indicators discriminated among different skills required for air combat. In summary, results from both the PCA and internal consistency computations suggest that the Gradesheet is a one-dimensional measurement instrument.

Discussion

It was hypothesized that the Gradesheet is sensitive to changes in air combat performance (a) over time in DMO training, (b) across team experience levels, and (c) among performance indicators. Results suggest that the Gradesheet did predict changes in team performance over a week of DMO training. For all indicators of 4-ship performance, the average team trajectory of estimated linear change on graded performance over missions was significantly positive. These results are largely consistent with predictions generated by SMEs on performance over missions. With the exception of Visual Lookout and Clear Avenue of Fire, the majority of SMEs predicted an effect of mission progression on 4-ship team performance. That is, if we disregard predictions SMEs made regarding experience by averaging performance across high and low experience levels, we find that the majority of SMEs judged that average performance of 4-ship teams increases over a DMO syllabus for all but two indicators. Further, that performance increases with time in DMO is what we expect given all the previous research on within-simulator learning.

However, the Gradesheet did not appear to be sensitive to differences in experience among teams on change in air combat performance over missions. Recall that past research found the Situation Awareness Rating Scale predictive of performance across different experience levels. Likewise, for 24 indicators found on the Gradesheet, the majority of SMEs that we surveyed in the current research predicted that change in average team performance over missions would be moderated by team F-16 experience. Specifically, for 17 indicators, more experienced teams were predicted by SMEs to perform better at the onset of training, but demonstrate less improvement over missions than less experienced teams. Results do not support these predictions. An experience by mission progression interaction on graded

performance wherein less experienced teams demonstrated greater improvement over missions than more experienced teams was observed for only one indicator: Egress – Separation (see Table 5), which was predicted by SMEs. For all other indicators, no effect of experience on performance over missions was observed. In fact, individual growth trajectories of teams on graded performance over missions were significantly variable for only six indicators, three of four Radar Mechanics indicators: Range Control, Azimuth Control, and Utilizing Correct Mode, as well as graded Gameplan Tactics, Detection to Commit, and Targeting. For none of these indicators did experience explain variability in slopes on graded performance among teams, even though this pattern of results was predicted by SMEs on the Radar Mechanics indicators and Gameplan Tactics. These results suggest that the Gradesheet is not sensitive to different experience levels among teams on trajectories of graded air combat performance.

Finally, the Gradesheet did not distinguish performance among indicators. From predictions generated by SMEs, we expected interactions between graded performance over missions and experience to differ by indicator. This was not observed. For the most part, results were consistent among indicators: First, graded performance was found to improve for all indicators. Recall that the majority of SMEs predicted no change in performance for two indicators: Visual Lookout and Clear Avenue of Fire. Second, although team experience was predicted to moderate graded performance over missions for 24 indicators, this result was observed for only one indicator. Third, for all indicators, graded 4-ship team performance was best estimated to improve linearly. Given that building block structure of DMO (crawl, walk, run), we expected that for some indicators of performance, improvement would be systematically curvilinear, with steep increases in performance observed at the beginning or end of the training week, depending upon the indicator, and the experience level of the team. This was not the case; adding a quadratic term to the model did little in explaining the nature of change. Finally, results of the principal components analysis suggest that most of the variance in graded performance among indicators can be explained by a single component. That this component represents anything more than general performance is unclear. In sum, the evidence suggests that the Gradesheet is not sensitive to performance differences among indicators.

Not only was graded performance not predictive of differences among experience levels and indicators, raising questions about the Gradesheet's predictive validity, but graded performance was also found to lack reliability for measuring air combat performance. Inter-grader reliability was low. Because SMEs were not consistent in grading performance using the Gradesheet, doubts about the Gradesheet's validity are amplified.

That graded performance lacked sensitivity to differences in performance among experience levels and indicators may be explained in several ways: First, the Gradesheet is not sensitive enough to capture differences in performance among experience levels and indicators. Second, the Gradesheet can detect such differences, but graders in the current investigation did not have the training necessary to use Gradesheet with this sensitivity. Third, the Gradesheet can distinguish differences in air combat performance among experience levels and by indicators, but in the current research, these effects were not present. We believe the third explanation is the least plausible. Previous research and predictions from SMEs in the current research suggest that differential effects of performance among experience levels and indicators should be observed. The more plausible explanation for the current findings are a combination of the first two

possibilities: Most likely the Gradesheet lacks sensitivity necessary to distinguish performance among experience levels and indicators; and SMEs did not have sufficient training to effectively use the Gradesheet.

Sensitivity of 4-ship Team Performance Measures

Results from the current research highlight the need for further development of process measures of air combat performance. We found very little evidence in the current research that the measurement system is sensitive to differences in performance among different indicators, or that process was measured at all. As noted several times, this measurement system has evolved from a line of research on situation awareness (Waag & Houck, 1995). Although the specific "behavioral elements" that have comprised the measure have changed several times over numerous iterations, the underlying content have remained unchanged. Indicators that have changed or been added to the Gradesheet over iterations are simply elaborations of concepts on previous iterations of the measure.

Even as the underlying content of the measurement system has remained unchanged over iterations, the scaling system that raters have used to assess performance has changed over time. The initial measure, the Situation Awareness Rating Scale, had a 6 point scale, with 1 indicating "acceptable", and 6 indicating "outstanding" (Waag & Houck, 1995). A subsequent streamlined version of the scale employed a 3 point scale, with 0, +, and – indicated average, above average, and below average mission performance (Crane, Robbins, & Bennett, 2000). The current measurement system uses a 5 point rating scale to assess performance, ranging from 0 to 4, with additional rating options of dangerous and not applicable to the current scenario. Even though the scale has varied in practice over different iterations of the measure, it is interesting to note that none of the findings from this research provide systematic evidence that performance varied by indicator type. Rather, results from this research have consistently been reported in the form of composite measures of air combat performance. And from this research, it is unclear whether a single indicator of general performance would not capture performance just as well as a composite of multiple indicators.

In generating subjective measures of air combat performance, a primary challenge has been developing measures that capture sufficient variability in performance. Capturing variability in performance was one of the motivating factors for changes made to the current Gradesheet over the years. Several methods have been used to address the challenge of capturing variability in rated performance. One of the earliest attempts to increase variability in graded performance was to simply increase the range of the scale (Payne, et al., 1976). This strategy was relatively unsuccessful, as scales with relatively small and large ranges resulted in rated performance clustered at the center of either scale. Researchers have also attempted to capture variability among indicators of air combat performance by developing measurement systems that assess specific behaviors rather than more general concepts, the rationale being that raters would be more reliable at judging specific rather than general performance. However, even this research has had limited success, especially at differentiating air combat performance among specific behavioral indicators. In the current research, researchers measured performance with behavioral indicators that were rated on a scale anchored by general performance objectives. Again, little variability in graded performance was observed in the current research;

grades were essentially clustered in the center of the scale. Because null effects are rarely published, one can easily imagine that other researchers have encountered similar problems in capturing variability among different indicators of air combat performance.

One explanation for clustering of rated air combat performance is based on experience SMEs have had grading performance at operational units. Many of the behavioral indicators included on the current measurement system are the same as or similar to items on Gradesheets used at operational units to measure air combat performance (Seaman, 1999). Most SME raters have had extensive history using these Gradesheets to assess performance. The legacy of these Gradesheets prescribes that instructors grade most performance as average, such that with the current Gradesheet most engagements would be graded as a 2 or 3. The consequences of grades lower than average can be high, including interruption of a career path as an air combat pilot. Outside of an operational environment and in a research context, it is possible that SMEs continue to rate most performance as average out of habit (Payne, 1982). This would explain why results from many research projects have found performance clustered at the center of rating scales.

Grader Training

The argument for carry over effects from SMEs experiences grading performance at operational units to SMEs experiences rating performance in the current research underscores the need for grader training. SME raters in the current research had no systematic training in how to employ the current measurement system. Such training could have potentially relegated questions on the degree to which SMEs used a sliding scale when measuring performance, wherein SMEs adjusted grading to account for increased scenario complexity over a week of DMO. Further, grader training could potentially illuminate and diminish inconsistencies in the way SMEs operationalize the rating scale for specific behavioral indicators. For example, what does a 1, 2, or 3 represent when grading Gameplan Tactics? From the current research, we have two sources of evidence suggesting variability among SMEs in their operational definitions for behavioral indicators: (a) for few indicators of performance were SMEs entirely consistent in their predictions of how performance changes over missions of a DMO syllabus, and (b) for all indicators low inter-grader reliability was observed on graded team performance. Presumably, if SMEs were better trained at using the current measurement system, then inter-grader reliability would have been much higher than was observed. However, grader training can only be effective if the measurement system allows raters to effectively discriminate differences in performance among different indicators of performance, and within a single indicator of performance. It is not clear that this is possible with the DMO Gradesheet.

Explaining Improvement in Graded Team Performance over Missions

For all indicators on the Gradesheet, average performance increased linearly across missions of a DMO syllabus. That results were consistent across all indicators suggests that the DMO Gradesheet may not be sensitive to changes in performance over missions. That is, increases in graded performance over missions may be explained by other factors. Ideally an evaluation of the Gradesheet would include SMEs who were blind to factors that could

potentially bias their grading of air combat performance. However, that was not the case in the current research, primarily because we were constrained in the current evaluation to use archival data collected at the AFRL in Mesa between August 2000 and December 2001. When SMEs graded 4-ship team performance during this time period, they had access to knowledge about pilots and situations that the pilots were in that potentially biased their grading. This knowledge could explain increases in graded performance over missions observed for all indicators.

Grader bias resulting from knowledge of participant characteristics. How plausible is it that knowledge SMEs had about pilots and teams influenced how they graded team performance? SMEs had access to and presumably knew multiple descriptive characteristics of the pilots and teams that they were grading: (a) the squadron they were from, (b) their rank, (c) their qualification level, and (d) the type and number of flying experiences they had. In some cases, SMEs had interacted with pilots before they arrived for training in Mesa, especially when pilots were participating in DMO for a second or third time, and thus had even more knowledge that could have contaminated how they graded performance. Further, from the first day of training on, SMEs knew the syllabi and related scenarios that teams would be flying, which is suggestive of the experience level of the team. In short, SMEs gain a great deal of knowledge about pilot teams before and during training that may potentially bias their grading. Even as SMEs were instructed to, and attempted to, disregard this information, it is possible that the information still had an unintentional impact on grading. What evidence is there that SME's knowledge of pilot characteristics biased graded performance? Because we asked SMEs to generate predictions about the effects of DMO on quality of performance for teams of more and less experience, we have some evidence to address this question. If we accept that team F-16 experience is a valid measure of experience, and that there was sufficient statistical power to capture the effect of experience on mean performance trajectories if there was one, then we can make the following argument: Because experience did not moderate the slope of graded performance trajectories over missions, it follows that any knowledge SMEs had of pilots' experience did not bias their judgments of performance over engagements and missions. For most indicators of performance, SMEs predicted that experience moderates performance over missions, but this was not the observed pattern of performance, suggesting that experience was not driving SMEs grading of performance over missions. However, if we consider graded performance at Mission 3, we do find that experience was related to performance for six indicators. Because we did not have SMEs generate predictions on performance at the onset of training, we have no information to assess whether SMEs expectations may be driving their grades at Mission 3.

Grader bias resulting from knowledge of the situation. What evidence is there that SME's situational knowledge contaminated their grading? When SMEs graded performance, they were continuously aware of how many DMO missions and scenarios that 4-ship teams had flown. So, as SMEs gained more knowledge of how skilled teams were, by observing their performance as they progressed through engagements and missions, it is not unlikely that they acquired expectations of how teams would perform in subsequent missions and engagements. It is possible that these expectations affected grading. Moreover, because most SMEs at the AFRL have much experience observing and judging 4-ship F-16 performance, they undoubtedly have a historical baseline to compare teams that they are currently grading to, which makes grader bias even more plausible. Furthermore, SMEs at the AFRL in Mesa are routinely asked to contribute

to the development of the DMO syllabus, so that training is as effective as possible. This is a requirement of their job. Thus, SMEs have a vested interest in demonstrating that DMO is effective. So, when SMEs are asked to grade performance in light of all the situational information they have, it is possible, if not probable, that this knowledge biases their grading, even if they are instructed to and attempt to disregard such information.

Besides knowledge of how team performance changed over measurement occasions, SMEs also had contextual knowledge during each measurement occasion that could have potentially biased their grading. When SMEs judged team performance during each scenario, it is possible that how they graded performance on some indicators influenced how they graded performance on other indicators. The Gradesheet is designed so that indicators are listed in an order that is notionally representative of order that cues corresponding to each indicator would be observed in the sequence of a scenario. Thus, graded performance among indicators is not independent. (But events in a scenario are not independent either.) For example, because Radar Mechanics indicators are serially positioned before Targeting on the Gradesheet, it is possible that graded team performance on Radar Mechanics indicators influences graded performance on Targeting. Moreover, in practice, the temporal nature of a scenario has little effect on grading because SMEs often did not grade team performance as a scenario plays out, but rather grading occurs after the scenario had been completed. That is, SMEs graded performance on each indicator in light of the outcomes of the scenario: kill ratio, exchange ratio, number of fratricides, etc… Thus, a highly plausible explanation for the current findings is that SMEs' knowledge of performance outcomes influenced graded performance on process indicators. Specifically, because the average slope of estimated trajectories of graded team performance over time was unexpectedly consistent among indicators, we question the construct validity of the Gradesheet - given the way it was used in the current research. Is graded performance on specific indicators a valid measure of performance for that indicator, or is it simply suggestive of the overall engagement grade? Evidence from the Principal Components Analysis (PCA) of the data suggests that one component explains most of the variability in graded performance among indicators; and that different types of indicators are not capturing any more or different information than other types of indicators. This is not a novel finding. Waag and Houck (1995) found that the majority of the variance among indicators in their measure of situation awareness could be explained by a single component, which they interpreted as a general composite measure of situation awareness? An alternative explanation is that the component they extracted is a composite measure of general performance rather than situation awareness.

In the research presented here, it is hard to discount the real threat that SMEs were biased by knowledge of the situation when grading performance, given the consistency among indicators in graded performance over missions, and among teams on specific indicators. None of the results in the present research argue against the possibility of grader bias from situational factors. Rather, they argue for it. Graded performance increased linearly and at a rather consistent rate over all indicators, with average estimated improvement among indicators of .082 grades per mission (*SD* = .012). Recall that SMEs hypothesized no change in performance for Visual Lookout and Clear Avenue of Fire, as well as an interaction between mission progression and experience for multiple indicators. Had SMEs been blind to situational factors when grading performance – mission count, performance outcome, and the like – it is plausible that no change in performance over missions would have been observed.

Future Directions: New Measures of Air combat Performance

Recently, research in collaboration with the AFRL in Mesa has focused on developing a new subjective measurement instrument. This instrument, SPOTLITE, is not only behaviorally based, but it is also behaviorally anchored (Schreiber, MacMillan, Carolan, & Sidor, 2002). That is, in SPOTLITE each behavioral indicator is related to air combat performance that is defined by specific behavioral criteria that anchor scores of the rating scale for the respective indicators. These behavioral indicators are tied to events that occur in the sequence of an air combat engagement. Mission Essential Competencies, as defined in Colegrove and Alliger (2002), serve as the underlying theoretical framework that is being used to define which knowledge, skills, and abilities are relevant to different phases of an-air combat engagement. Although the content of MECs overlaps with many of the indicators listed in the DMO Gradesheet, MECs expand on these concepts by linking them to specific phases of air combat missions with greater specificity. Thus, the system comes closer to meeting the definitional requirement of air combat measurement that was posited by Waag, Pierce, and Fessler (1987). Preliminary research results suggest that SPOTLITE is more sensitive than past measures to differences in air combat performance along multiple dimensions.

In addition to generating a better subjective measurement system of air combat performance, the AFRL in Mesa has recently developed an objective measurement system, the Performance Evaluation Tracking System (PETS) (Schreiber, Watz, Bennett, & Portrey, 2003). PETS is a software tool that enables multi-platform, multi-level measurement ability at the F-16 individual and team level in distributed environments. Approximate one million data points per minute are collected and organized into several file formats. These data points include demographical, positional, and operational variables that are used to generate files with summary measures for shot and engagements. PETS measures performance that is not easily assessed with reliability by subjective rating systems. For example, PETS can compute how far an entity has penetrated an aircraft's weapons engagement zone (WEZ) by using instantaneous entity state vector information and weapons tables. Thus, in current research at the AFRL, subjective and objective measures are being developed to complement each other. Tasks that are not easily defined using PETS are being assessed with subjective measures, and tasks that are better quantified with PETS are not being explored in detail with subjective measures.

Conclusion

The primary objective of the current research was to evaluate the effectiveness of DMO Gradesheet at measuring air combat skill. Results suggest that the Gradesheet lacked sufficient reliability, validity, and sensitivity when used to measure 4-ship team performance in the simulated DMO air combat environment. Inter-grader reliability estimates were low, suggesting that given limited training, raters cannot consistently use the Gradesheet to measure performance. As expected, trajectories of graded team performance increased over missions. However, because the pattern of graded performance over missions was largely consistent across all indicators, it is possible that graders were unintentionally biased by knowledge they had about pilots and the missions they were flying. We predicted differences in performance over missions

among team experience levels and indicators. These results were not observed, suggesting that the Gradesheet is not sensitive enough to discriminate different aspects of air combat performance at any experience level. The findings suggest that a single indicator of general performance, similar to an Overall Engagement Grade, would have been just as effective at explaining performance as the 40 indicators within the Gradesheet. Finally, because graders were not blind when grading performance, even the validity of the Gradesheet as a gross measure of air combat performance is of question. To better measure air combat performance in the future, subjective measurement instruments under development in collaboration with the AFRL in Mesa will distinguish performance within indicators of performance, thus increasing the sensitivity that air combat is measured. Because subjective measures will be based on MECS, the validity with which air combat is measured will be enhanced as well. Further, these subjective measures will complement objective measures that are being collected directly from the simulated DMO environment.

References

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology. 8* (3), 223-242.

Bennett, W., Schreiber, B. T., & Andrews, D. H. (2002). Developing competency-based methods for near-real-time air combat problem solving assessment. *Computers in Human Behavior, 18,* 773-782.

Breck, F. H., & Miller, D. C. (1991). *Aircrew performance measurement in the air combat maneuvering domain: A critical review of the literature.* (AL-TR-1991-0042/AD B158404). Williams Air Force Base, AZ: Armstrong Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Colegrove, C. M., & Alliger, G. M. (2002). *Mission essential competencies: Defining combat mission requirements in a novel way.* Paper presented at the NATO SAS-038 Research and Technology Organization, Brussels, Belgium.

Crane, P., Robbins, R., & Bennett, W. (2000). Using distributed mission training to augment flight lead upgrade training. In, *Proceedings of Interservice/Industry Training Systems and Education Conference (I/ITSEC)*. (pp. 1175-1184). Orlando, FL: National Security Industrial Association.

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1993). *An information processing classification of beyond-visual-range air intercepts.* (AL/HR-TR-1993-0061, AD A266 927). Williams Air Force Base, AZ: Armstrong Laboratory, Human Resources Directorate, Aircrew Training Research Division. May 1993.

Kelly, M. J. (1988). Performance measurement during simulated air-to-air combat. *Human Factors*, 30, 495-506.

Kenny, D. A., Bolger, N., & Kashy, D. A. (2002). Traditional methods for estimating multilevel models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data*. (pp. 1-24). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel models.* London: Sage Publications.

McGuiness, J., Bouwman, J. H., & Puig, J. A. (1982). Effectiveness evaluation for air combat training. In, *Proceedings of the 4th Interservice/Industry Training Equipment Conference.* (pp. 391-396). Washington D.C.: National Security Industrial Association, 16-18 November 1982.

Payne, T. A. (1982).*Conducting studies of transfer of learning: A practical guide.* (AFHRL-TR-81-25). Williams Air Force Base, AZ: Operations Training Division, Air Force Human Resources Laboratory.

Payne, T. A., Hirsch, D. L., Semple, C. A., Farmer, J. R., Spring, W. G., Sanders, M. S., Wimer, C. A., Carter, V. E., & Hu, A. (1976). *Experiments to evaluate advanced flight simulation in air combat pilot training: Vol 1. Transfer of learning experiment.* Huntsville, AL: Northrop Corporation.

Schreiber, B. T., Watz, E., & Bennett, W. (2003). Objective human performance measurement in a distributed environment: Tomorrow's needs. In, *Proceedings of Interservice/Industry Training Systems and Education Conference (I/ITSEC).* Orlando, FL: National Security Industrial Association. 03-BRIM-062.

Schreiber, B. T., Watz, E., Bennett, W., & Portrey, A. M. (2003). Development of a distributed mission training automated performance tracking system. In, *Proceedings of Behavior Representation in Modeling and Simulation Conference,* Scottsdale, AZ.

Schreiber, B. T., MacMillan, J., Carolan, T. F., & Sidor, G. (2002). *Evaluating the effectiveness of distributed mission training using 'traditional' and innovative metrics of success.* Paper presented at the NATO SAS-038 Working Group Meeting, Brussels, Belgium.

Seaman, K. A. (1999). *Improving F-15C air combat training with distributed mission training (DMO) advanced simulation.* (AU/ACSC/183/1999-04). Maxwell Air Force Base, AL: Air Command and Staff College Air University.

Tabachnick, B. G., and Fidell, L. S. (2001). *Using multivariate statistics.* (4th Ed.). Boston: Allyn & Bacon.

TAC Regulation 50-31. (1983). *Training records and performance evaluations in formal flying training programs.* Langley AFB, VA: HQ TAC.

Waag, W. L., & Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, Space, and Environmental Medicine*, *65* (5, Suppl.), A13-A19.

Waag, W. L., & Houck, M. R. (1995). Development of criterion measures of situation awareness for use in operational fighter squadrons. In AGARD Conference Proceedings 575 *Situation Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD-CP-575). (pp. 8-1-8-8). Neuilly-Sur-Seine, France: Advisory Group for Aerospace Research & Development.

Waag, W. L., Houck, M. R., Greschke, D. A., & Raspotnik, W. B. (1995). Use of multiship simulation as a tool for measuring and training situation awareness. In AGARD Conference Proceedings 575 *Situation Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD-CP-575). (pp. 20-1-20-8). Neuilly-Sur-Seine, France: Advisory Group for Aerospace Research & Development.

Waag, W. L., Pierce, B. J., & Fessler, S. (1987). *Performance measurement requirements for tactical aircrew training.* (AFHRL-TR-86-62). Williams Air Force Base, AZ: Armstrong Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Waag, W. L., Raspotnik, W. B., & Leeds, J. L. (1992). *Development of a composite measure for predicting engagement outcome during air combat maneuvering.* (AL-TR-1992-0002, AD A252344). Williams Air Force Base, AZ: Armstrong Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Appendix A
Predicted DMO Performance by SMEs

Introduction

In the history of research on within-simulator learning of air combat skill, little evidence has been collected assessing how specific skills necessary for successful air-to-air combat change over missions of a training syllabus. For the most part, process measures of performance have been composites that summarize overall performance on a given engagement. When performance data has been collected on specific processes, analyses have focused on performance aggregated from all indicators (Seaman, 1999). To get a better understanding of how different processes may change over missions of a DMO syllabus, we asked F-16 Subject Matter Experts (SMEs) to use their expertise to generate predictions about how Mesa DMO affects performance of 4-ship teams as they progress through a week of DMO. Moreover, SMEs were asked if the impact of DMO on performance is different for teams that have more or less F-16 experience among the 4-ship, because air combat flight experience has been found to consistently predict outcomes of air combat engagements (Waag & Houck, 1995).

Method

Participants

Six SMEs were recruited to generate these predictions based on their extensive experience and knowledge of DMO at the AFRL in Mesa. On average, SMEs had more than 2000 hours of F-16 flight time in careers that exceeded ten years.

Questionnaire

To generate predictions, SMEs were given a questionnaire that provided them with eight graphs depicting different performance patterns that might occur for 4-ship teams with more or less flight experience (see the full questionnaire in Appendix B). Although total F-16 experience of pilots within teams varies continuously, we decided to make experience a dichotomous variable in the questionnaire to make interactions of mission progression by experience easier to depict and interpret graphically. For this exercise, more and less experienced teams were defined as having respectively 5000 and 2500 combined F-16 hours among the 4 pilots. These hours were based on the experience of the 4-ship teams whose data were used to evaluate the DMO Gradesheet, as described in the Method section below. Graphs had mission number on the abscissa and quality of performance on the ordinate, and depicted linear growth trajectories for more and less experienced teams. Patterns 1 through 4 depicted different possible main effects of mission progression and/or experience on quality of performance. As can be seen, all effects of experience had more experienced teams performing better than less experienced teams during any given mission. Patterns 5 through 8 depict possible interactions between mission progression and experience on quality of performance. Patterns 5 and 6 show improvement for both more and less experienced teams, but vary in the rates of improvement between the types of teams. Finally, Patterns 7 and 8 depicted performance gain for only one level of experience, and no gain for the other experience level:

In the questionnaire (Appendix B), SMEs were asked to consider in turn each of 40 indicators of 4-ship team performance, and choose the pattern for each indicator that best represented their judgment of how quality of team performance progresses during a DMO training week for more and less experienced teams. The 40 indicators of performance came from the DMO Gradesheet, and are defined in Appendix B. In making their judgments, SMEs were given several instructions designed to limit assumptions that they could make when completing the questionnaire. First, they were asked to base their judgments on their beliefs about how DMO affects different indicators of team performance, rather than their knowledge of what past research has said about how DMO affects performance. Second, SMEs were instructed to make their judgments independent of the many factors that can impact team performance in Mesa DMO. So, for example, although teams fly many different scenarios in a given training week, SMEs were instructed to make their judgments independent of scenario, scenario complexity, and syllabus type. Finally, if SMEs found that the patterns provided did not sufficiently capture the pattern of performance they expected, then they were given the option of drawing a graph depicting the pattern that they judged best described performance. This option was provided because we believed that for some indicators expected growth trajectories would be curvilinear, which was an added complexity that we did not want to incorporate into the prediction generation task.

Results

Frequencies of predicted patterns of performance generated by SMEs on all indicators are presented in Table A1. As can be seen, SMEs judgments of quality of performance varied both within and among different indicators. Nonetheless, structure does emerge from hypotheses generated by the six SMEs. First, for 37 of 40 indicators, if we average over experience levels, the majority of SMEs judged that quality of performance increases over missions of a DMO syllabus, which is depicted in all patterns except 1 and 3. That is, only on 2 indicators did the majority of SMEs predict no change in performance across missions: Visual Lookout, and Clear Avenue of Fire. Second, for 24 indicators the majority of SMEs judged that experience moderates change in performance over missions, as depicted in patterns 5 through 8. For 9 other indicators the majority of SMEs judged that experience does not predict differential changes in performance across missions, as depicted in patterns 1 through 4. Third, when experience was expected to moderate change in performance over missions, pattern 5 was judged to best represent expected performance by the majority of the SMEs for 14 indicators. Pattern 5 depicts improvement in performance across missions for both more and less experienced teams, with less experienced teams improving at a faster rate than more experienced teams. Pattern 5 was the only pattern selected by four or more SMEs on any given indicator, with one exception.

***Table A1*. Frequency of predicted pattern of performance from six SMEs on forty indicators of 4-ship team performance.**

	Predicted Pattern of Performance							
Indicator	1	2	3	4	5	6	7	8
Radar Mechanics:								
El Strobe Control				1	5			
Range Control			1	1	3	1		
Azimuth Control					5		1	
Utilizing Correct Mode	1	1		1	2		1	
Gameplan:								
Tactics		1		3	1		1	
Execution				2	4			
Adjusting Plan On-the-Fly				1		4	1	
Tactical Intercepts:								
Formation		1		2	1		1	1
Detection – Commit	1			2	3			
Targeting				2	3		1	
Sorting				2	4			
BVR Launch and Leave				1	4	1		
BVR Launch and React					3	3		
Intercept Geometry				3	1	1	1	
Low Altitude Intercepts			1	3	2			
AAMD:								
RMD				2	2		1	1
IRCM			2	3			1	
Chaff – Flares		1		2	3			
Communications:								
3-1 Communication		1		1	4			
Radio Discipline				2	4			
GCI Interference				2	4			
Additional Indicators								
Engagement Decision					3	3		
Spike Awareness		1		3	2			
E F & N Pole			1	2	3			
Egress – Separation				1	4	1		
Contracts				2	3		1	
ROE Adherence	1			1	2		2	
ID Adherence	1	2		1	1		1	
Post Merge Maneuvering			2		3	1		
Mutual Support		1		1	2	1	1	
Visual Lookout	1	1	3	1				
Weapons Employment				2	4			
Clear Avenue of Fire	1	1	3	1				
Fuel Management	1	1	2		1		1	
Flight Discipline		3		1	2			
Situation Awareness					6			
Judgment				1	5			
Flight Leadership – Conduct				1	4	1		
Briefed Objectives Fulfilled				3	2	1		
Overall Engagement Grade		1			5			

Appendix B
DMO Performance Predictions Questionnaire

Instructions

The Training System Technology Team is in the process of generating several new research studies, and would appreciate your assistance. The purpose of this exercise is to use your expertise and tactical knowledge to generate hypotheses about how you think Mesa DMT affects performance of 4-ship teams as they progress across missions during a training week. To this end, please:

1. Familiarize yourself with the skill definitions on pages 2 and 3. Obviously these skills could be defined a number of ways, but for this exercise, please use the definitions provided.

2. Familiarize yourself with the charts on page 4. The charts depict different patterns of quality of performance for pilot teams of more (blue lines) and less (pink lines) experience as they progress across missions from Monday afternoon to Friday. For this exercise, assume that more experienced teams have approximately 5000 combined F-16 hours among the 4 pilots, while a less experienced teams have approximately 2500 combined F-16 hours.

3. Complete the tables on pages 5 and 6 in light of the skill definitions on pages 2 and 3 and the eight patterns of performance on page 4. For each indicator of performance listed in the tables, write the number corresponding to the one pattern from page 4 that best represents your judgment of how quality of team performance will progress during a training week for more and less experienced teams.

   Note:

   - ⇨ Base your judgments on your beliefs about how DMT affects different indicators of team performance, rather than your knowledge of what past research has said about how DMT affects performance.
   - ⇨ Although many factors can impact team performance in Mesa DMT, for this exercise, make your judgments independent of these factors. For example, although teams fly many different scenarios in a given training week, make your judgments independent of scenario, scenario complexity, syllabus type, etc...
   - ⇨ If you find that an expected pattern of performance for a given indicator is not described by options 1-8, please draw a graph depicting the pattern that you judge best describes performance on the back of pages 5 and 6.

For questions, please contact Mike Krusmark at ext 465.

Thank you for your time and assistance.

Skill Definitions

Radar Mechanics:

El Strobe Control – Did the pilots adjust elevation control to cover assigned airspace and/or bracket targets' altitudes?

Range Control – Did the pilots have the appropriate radar range selected and did they bump down ranges accordingly?

Azimuth Control – Did the pilots have radar in correct azimuth at specific times (i.e. 60, 30, or 10 degree sweep)?

Utilizing Correct Mode – Did the pilots use the right radar mode for radar/ threat presentation (i.e. RWS, TWS, and ACM modes)?

Gameplan:

Tactics – As a result of the briefing, how well did the pilots handle/ attack a specific threat presentation?

Execution – How well did the pilots handle threats?

Adjusting Plan On-the-Fly – If gameplan broke down, how well did the FL improvise to handle threats?

Tactical Intercepts:

Formation – Did the pilots fly in the correct position and maintain visual mutual support?

Detection / Commit – (relates to Radar Mech.) Did the pilots locate the targets and pursue?

Targeting – Did all pilots/ elements correctly target/ monitor briefed or assigned groups?

Sorting – Did all pilots correctly sort and/or radar lock assigned contacts within their assigned group?

BVR (beyond-visual-range) Launch and Leave – Did the pilots shoot and leave at the appropriate range (Pitbull) – usually outside of minimum abort range.

BVR (beyond-visual-range) Launch and React – Did the pilots shoot and react (Notch) when inside of MAR – usually with a 'spike'.

Intercept Geometry – Was the correct radar geometry used to perform the intercept? Did the pilots notice aspect changes and maneuver accordingly?

Low Altitude Intercepts – Was the correct intercept flown against low targets (altitude separation)?

Engagement Decision – How correct was the decision to engage targets? Was adequate range available, 'spiked' or 'naked', delouse criteria met, low vs. high risk area, etc?

Spike Awareness – Did the pilots know when they were spiked and react correctly?

E/F/N Pole – Did the pilots correctly perform appropriate maneuvering?

Egress / Separation – Was the decision to egress/ separate timely and effective?

AAMD:

RMD – Did the pilots perform correct and effective defense against radar missile threat?

IRCM – Did the pilots use correct and effective countermeasures against IR missile threat?

Chaff / Flares – Were chaff/ flares dispensed at the appropriate times?

Contracts – Did the pilots adhere to briefed contracts/ responsibilities?

ROE Adherence – Did the pilots follow/ adhere to briefed rules of engagement?

ID Adherence – Did the pilots follow/ adhere to briefed identifications requirements of bogey/ spades/ hostile aircraft?

Post Merge Maneuvering – Did the pilots maneuver aircraft correctly after merging with targets? Did they keep sight of bandits and perform effective ACM?

Mutual Support – Did the pilots provide effective mutual support to each other or each element – usually implies maintaining a visual within the element or positional situation awareness between elements.

Visual Lookout – Did the pilots look outside the cockpit to pick up each other or targets visually?

Weapons Employment – Were all missile shots taken between maximum and minimum ranges, and were the correct number of shots taken?

Clear Avenue of Fire – Did the pilots hold shots if a friendly aircraft was in line of fire?

Communication:

3-1 Comm – Did the pilots use correct 3-1 brevity codes and terms?
Radio Discipline – Did the pilots adhere to sound radio discipline and omit extraneous chatter?
GCI Interface – How well did AWACS and the pilots communicate with each other?

Fuel Management – Were the pilots aware of their fuel status and act accordingly?

Flight Discipline – Did the pilots adhere to briefed procedures and designated responsibilities?

Situation Awareness – How was overall SA in regards to each other and hostile groups?

Judgment – Were sound decisions made and were they made in a timely manner?

Flight Leadership / Conduct – How well did the flight lead control the flight?

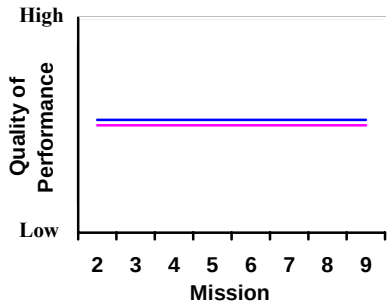
Briefed Objectives Fulfilled – Were the desired goals achieved?

Overall Engagement Grade – What was the team's overall performance?

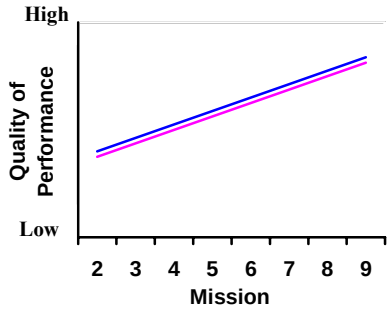
Possible Training Effects

More Experienced Teams Less Experienced Teams

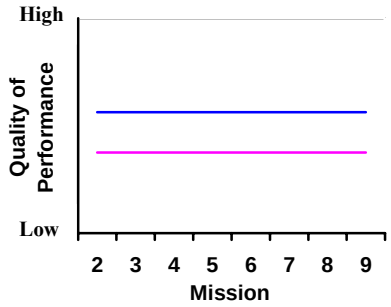
Pattern 1. No initial difference in quality of performance between teams of more and less experience, with no change in performance over missions for either type of team.



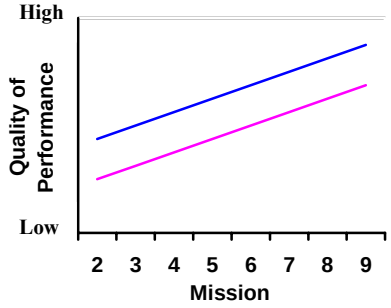
Pattern 2. No initial difference in quality of performance between teams of more and less experience, with performance improving at the same rate across missions for both types of team.



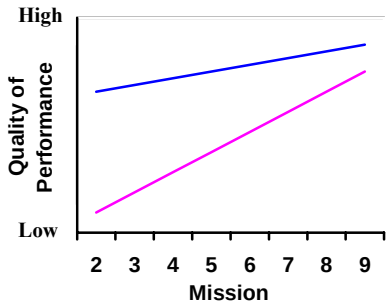
Pattern 3. Initially high experience teams perform better than low experience teams, with no change in quality of performance over missions for either type of team.



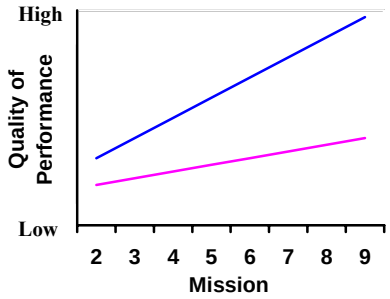
Pattern 4. Initially high experience teams perform better than low experience teams, with quality of performance improving at the same rate across missions for both types of team.



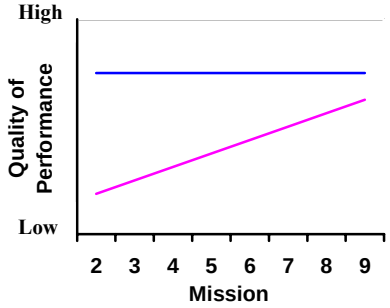
Pattern 5. Quality of performance improves across missions at a faster rate for low experience teams than for high experience teams.



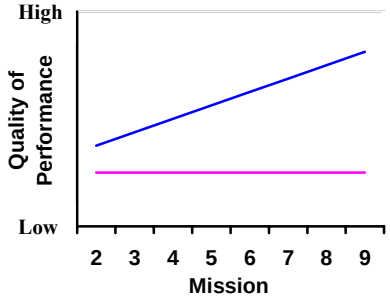
Pattern 6. Quality of performance improves across missions at a faster rate for high experience groups than for low experience teams.



Pattern 7. Quality of performance gain across missions is seen only for low experience teams.



Pattern 8. Quality of performance gain across missions is seen only for high experience teams.



For each process indicator of performance listed below, write the number that corresponds to the pattern from page 4 that best represents your judgment of how quality of performance progresses during DMT from Monday afternoon to Friday for more and less experienced 4-ship teams.

	Performance Indicator	Predicted Pattern of Performance (#)	Comments
1	Radar Mechanics: El Strobe Control		
2	Radar Mechanics: Range Control		
3	Radar Mechanics: Azimuth Control		
4	Radar Mechanics: Utilizing Correct Mode		
5	Gameplan: Tactics		
6	Gameplan: Execution		
7	Gameplan: Adjusting Plan On-the-Fly		
8	Tactical Intercepts: Formation		
9	Tactical Intercepts: Detection – Commit		
10	Tactical Intercepts: Targeting		
11	Tactical Intercepts: Sorting		
12	Tactical Intercepts: BVR Launch and Leave		
13	Tactical Intercepts: BVR Launch and React		
14	Tactical Intercepts: Intercept Geometry		
15	Tactical Intercepts: Low Altitude Intercepts		
16	Engagement Decision		
17	Spike Awareness		
18	E F & N Pole		
19	Egress – Separation		
20	AAMD: RMD		
21	AAMD: IRCM		
22	AAMD: Chaff – Flares		
23	Contracts		
24	ROE Adherence		
25	ID Adherence		
26	Post Merge Maneuvering		
27	Mutual Support		
28	Visual Lookout		
29	Weapons Employment		
30	Clear Avenue of Fire		
31	Communications: 3-1 Communication		
32	Communications: Radio Discipline		
33	Communications: GCI Interface		
34	Fuel Management		
35	Flight Discipline		
36	Situation Awareness		
37	Judgment		
38	Flight Leadership – Conduct		
39	Briefed Objectives Fulfilled		
40	Overall Performance		

For each outcome indicator of performance listed below, write the number that corresponds to the pattern from page 4 that best represents your judgment of how quality of performance progresses during DMT for more and less experienced 4-ship teams.

	Outcome Indicator	Predicted Pattern of Performance (#)	Comments
1	Proportion of Threats Killed		
2	Proportion of Viper Shots resulting in a Kill		
	For the following indicators, high quality of performance represents smaller numbers		
3	Proportion of Vipers Killed (high quality of performance indicates fewer vipers killed)		
4	Number of Viper Fratricides (high quality of performance indicates fewer fratricides)		
5	Proportion of Threat Shots resulting in a Kill		
6	Proportion of Enemy Fighters alive at end of flight		
7	Proportion of Enemy Bombers Alive at end of flight		
8	Total Time Vipers were in Minimum Abort Range		
9	Total Time Vipers were in Maximum Out Range		
10	Closest 2D Range any Viper came to a Threat		

Appendix C
Gradesheet as Administered

Research ID Number:

_____

Rater ID Number:

Mission Evaluation Sheet

(Please rate the debrief using the following criteria)

Grading Criteria:

N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge
Grade 1: Performance is safe, but indicates limited proficiency.
Makes errors of omission of commission
Grade 2: Performance is essentially correct. Recognizes and corrects errors
Grade 3: Performance is correct, efficient, skillful and without hesitation
Grade 4: Performance reflects an unusually high degree of ability

Brief

Briefing Task	Grade	Notes
1. Mission Preparation	NA D 0 1 2 3 4	
a. Developing Plan	NA D 0 1 2 3 4	
2. Briefing		
a. Organization	NA D 0 1 2 3 4	
b. Content	NA D 0 1 2 3 4	
c. Delivery	NA D 0 1 2 3 4	
d. Instructional Ability	NA D 0 1 2 3 4	
3. Systems Knowledge	NA D 0 1 2 3 4	
4. Overall quality of brief	NA D 0 1 2 3 4	

Mission Evaluation Sheet
(please rate the debrief using the following criteria)

Grading Criteria:

N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge
Grade 1: Performance is safe, but indicates limited proficiency.
Makes errors of omission of commission
Grade 2: Performance is essentially correct.
Recognizes
and corrects errors
Grade 3: Performance is correct, efficient, skillful and without hesitation
Grade 4: Performance reflects an unusually high degree of ability

Debrief

Debriefing Task	Grade	Notes
1. Debriefing		
a. Organization	NA D 0 1 2 3 4	
b. Reconstruction	NA D 0 1 2 3 4	
c. Delivery	NA D 0 1 2 3 4	
d. Analysis	NA D 0 1 2 3 4	
e. Instructional Ability	NA D 0 1 2 3 4	
2. ID Adherence	NA D 0 1 2 3 4	
3. Flight leadership	NA D 0 1 2 3 4	
4. Mission Objectives Accomplished	NA D 0 1 2 3 4	
5. Overall quality of debrief	NA D 0 1 2 3 4	

Research Gradesheet

Team: _____ Rater ID (last four): _____ Pilot ID Number (5 Digit ID): _____

SCENARIO ID	Additional threats presented	Level of difficulty	Engagement Number
			1 2 3 4 5 6 7 8

Grading Criteria:

N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge.
Grade 1: Performance is safe, but indicates limited proficiency. Makes errors of omission or commission
Grade 2: Performance is essentially correct. Recognizes and corrects errors.
Grade 3: Performance is correct, efficient, skillful and without hesitation.
Grade 4: Performance reflects an unusually high degree of ability

Engagement Task	Grade	Notes
1. Radar Mechanics		
a. El Strobe Control	NA D 0 1 2 3 4	
b. Range Control	NA D 0 1 2 3 4	
c. Azimuth Control	NA D 0 1 2 3 4	
d. Utilizing correct mode	NA D 0 1 2 3 4	
2. Game plan / Tactics	NA D 0 1 2 3 4	
a. Execution	NA D 0 1 2 3 4	
b. Adjusting Plan On-The-Fly	NA D 0 1 2 3 4	
3. Tactical Intercepts		
a. Formation	NA D 0 1 2 3 4	
b. Detection / Commit	NA D 0 1 2 3 4	
c. Targeting	NA D 0 1 2 3 4	
d. Sorting	NA D 0 1 2 3 4	
e. BVR launch and leave	NA D 0 1 2 3 4	
f. BVR launch and react	NA D 0 1 2 3 4	
g. Intercept Geometry	NA D 0 1 2 3 4	
h. Low altitude intercepts	NA D 0 1 2 3 4	
4. Engagement Decision	NA D 0 1 2 3 4	
5. Spike Awareness	NA D 0 1 2 3 4	
6. E/F/N Pole	NA D 0 1 2 3 4	
7. Egress / Separation	NA D 0 1 2 3 4	

Grading Criteria:

N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge.
Grade 1: Performance is safe, but indicates limited proficiency. Makes errors of omission or commission
Grade 2: Performance is essentially correct. Recognizes and corrects errors.
Grade 3: Performance is correct, efficient, skillful and without hesitation.
Grade 4: Performance reflects an unusually high degree of ability

Engagement Task	Grade	Notes
8. AAMD		
a. RMD	NA D 0 1 2 3 4	
b. IRCM	NA D 0 1 2 3 4	
c. Chaff / Flares	NA D 0 1 2 3 4	
9. Contracts	NA D 0 1 2 3 4	
10. ROE Adherence	NA D 0 1 2 3 4	
11. ID Adherence	NA D 0 1 2 3 4	
12. Post Merge Maneuvering	NA D 0 1 2 3 4	
13. Mutual Support	NA D 0 1 2 3 4	
14. Visual lookout	NA D 0 1 2 3 4	
15. Weapons Employment	NA D 0 1 2 3 4	
16. Clear Avenue of Fire	NA D 0 1 2 3 4	
17. Communication		
a. 3-1 Comm	NA D 0 1 2 3 4	
b. Radio Discipline	NA D 0 1 2 3 4	
c. GCI Interface	NA D 0 1 2 3 4	
18. Fuel Management	NA D 0 1 2 3 4	
19. Flight Discipline	NA D 0 1 2 3 4	
20. Situation Awareness	NA D 0 1 2 3 4	
21. Judgment	NA D 0 1 2 3 4	
22. Flight Leadership/ Conduct	NA D 0 1 2 3 4	
23. Briefed Objectives Fulfilled	NA D 0 1 2 3 4	

Objective Measures: Team Performance Statistics

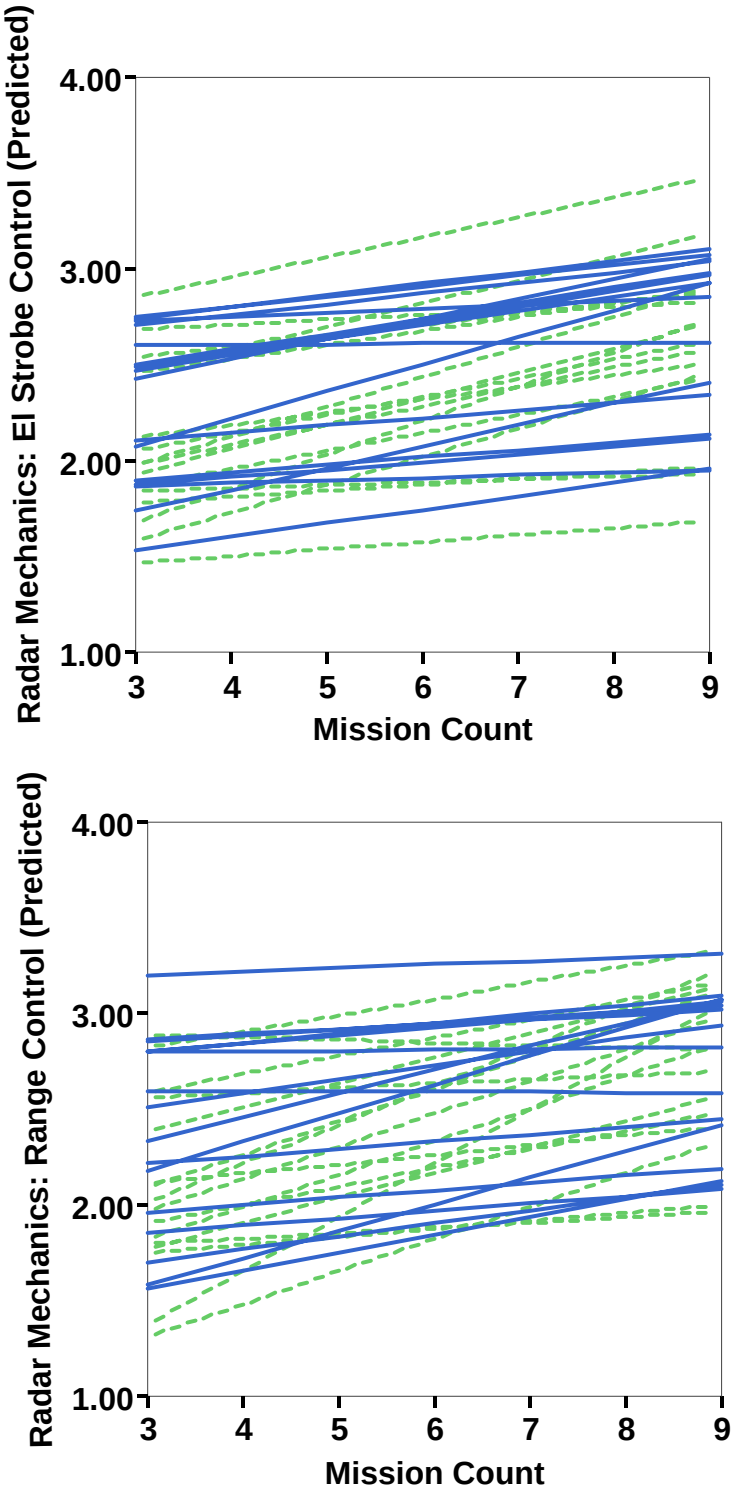
Viper	AIM-120	AIM-9	Gun	Number of Kills	# Invalid	Explanation
1	1 2 3 4 5 6	1 2 3 4	1 2	1 2 3 4 5 6 7 8		
2	1 2 3 4 5 6	1 2 3 4	1 2	1 2 3 4 5 6 7 8		
3	1 2 3 4 5 6	1 2 3 4	1 2	1 2 3 4 5 6 7 8		
4	1 2 3 4 5 6	1 2 3 4	1 2	1 2 3 4 5 6 7 8		
TOTALS						

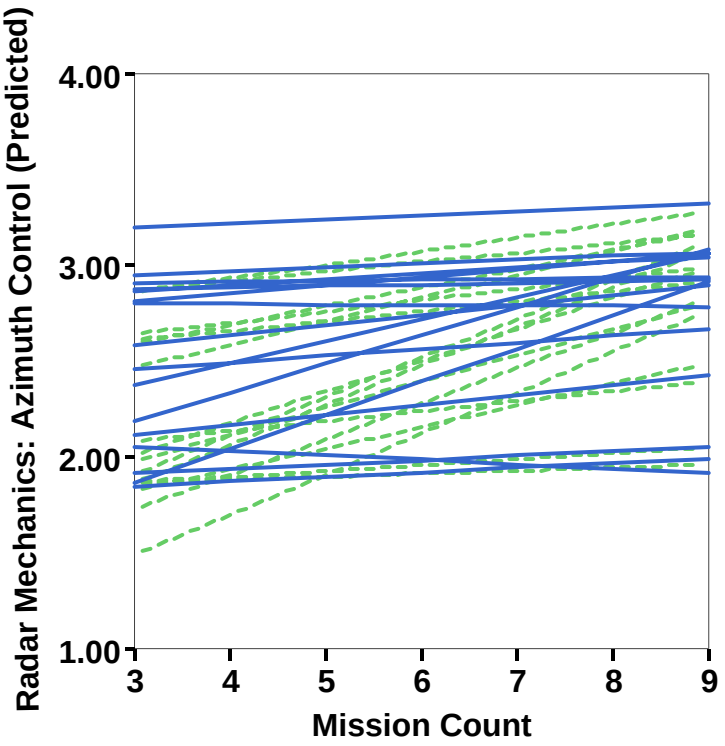
Survivors _____ Morts _____ Frats _____

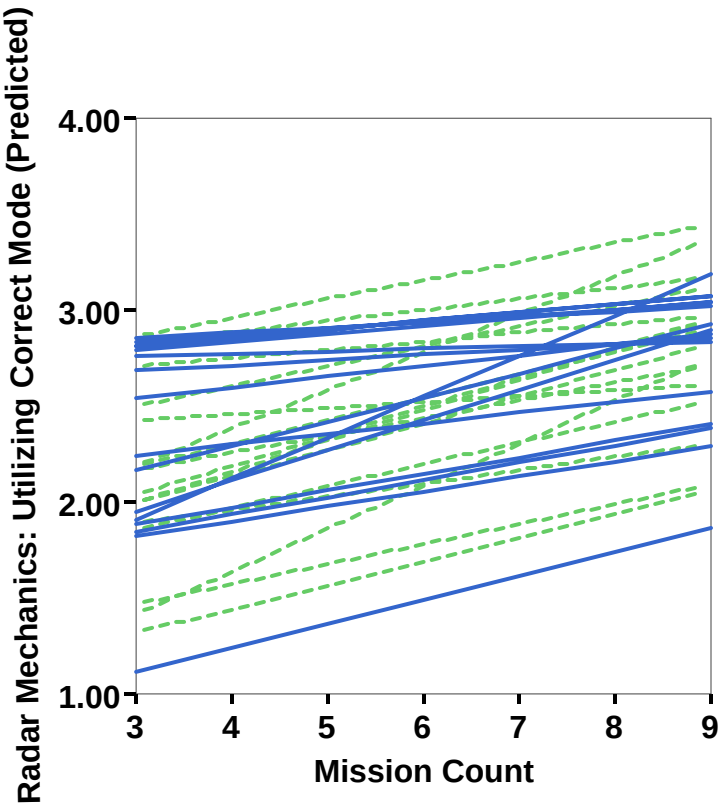
Overall Engagement Grade: 0 1 2 3 4

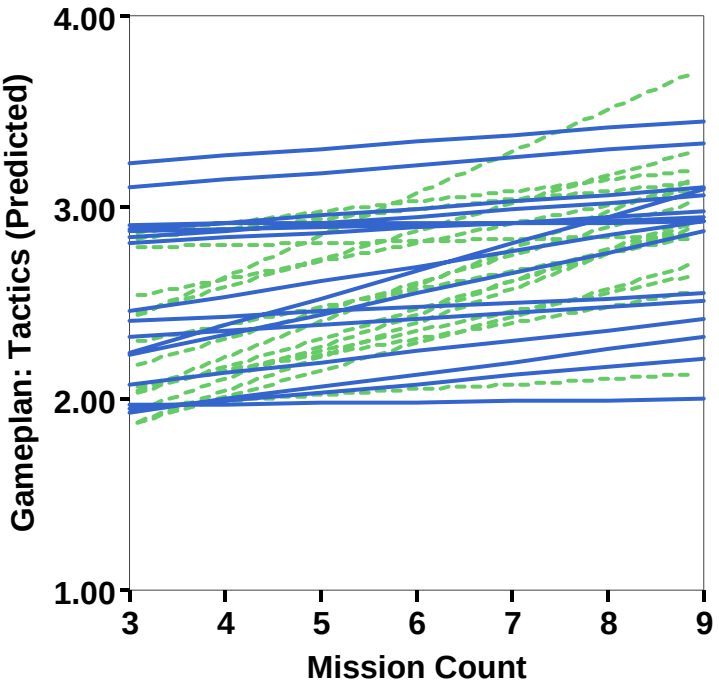
Appendix D
Figures

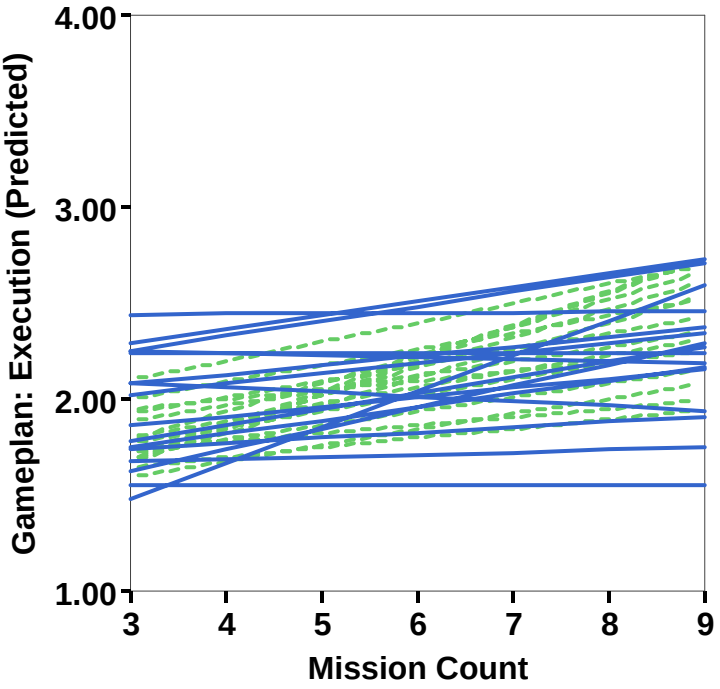
Figure 2. **Estimated linear change trajectories of thirty-two 4-ship teams on graded performance plotted separately by indicator. Solid blue lines and dashed green lines depict estimated performance of teams with more and less 4-ship F-16 experience than the median team respectively.**

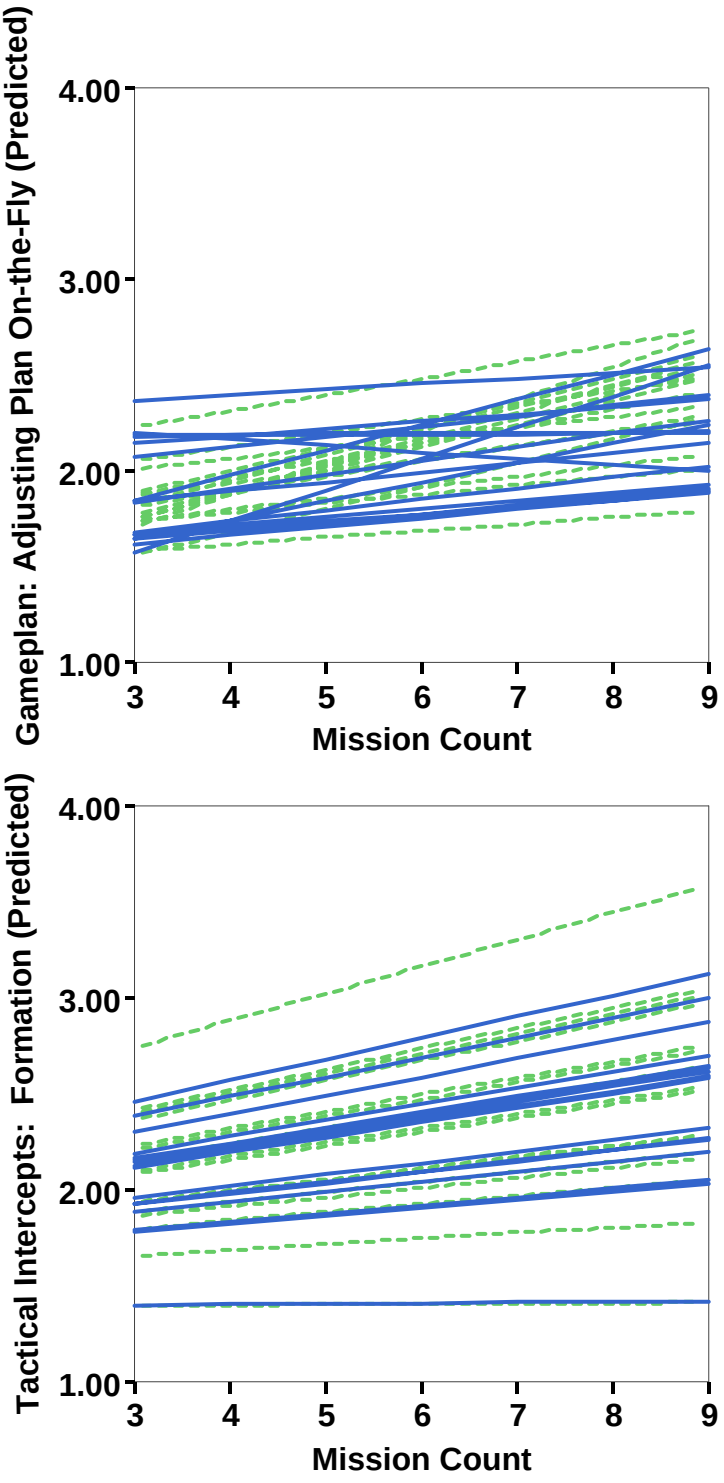


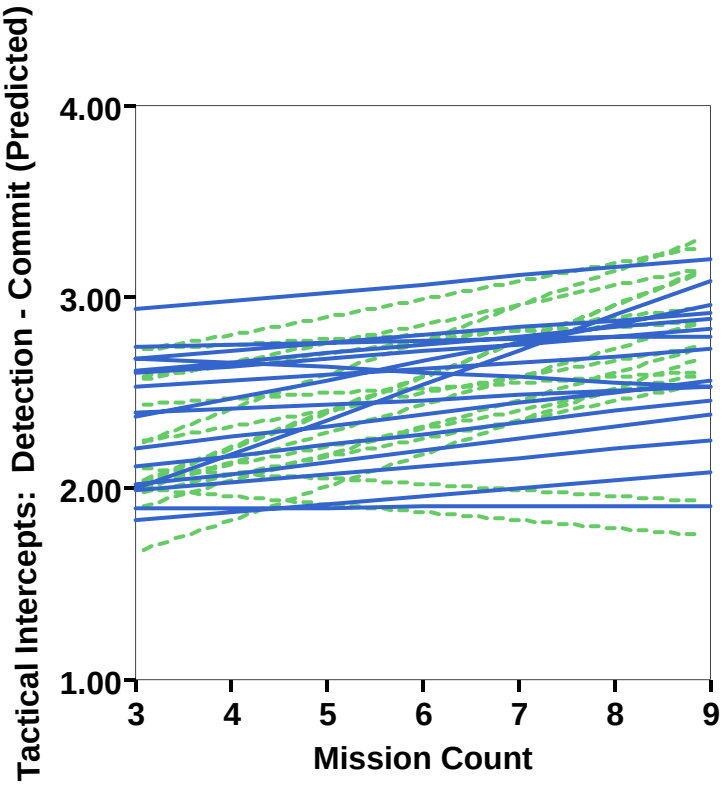

Radar Mechanics: Azimuth Control (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

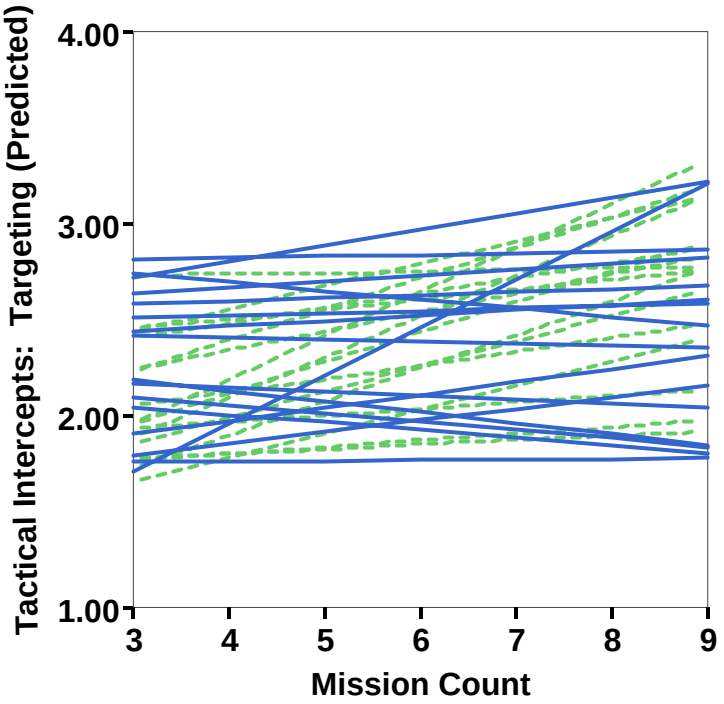

Radar Mechanics: Utilizing Correct Mode (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

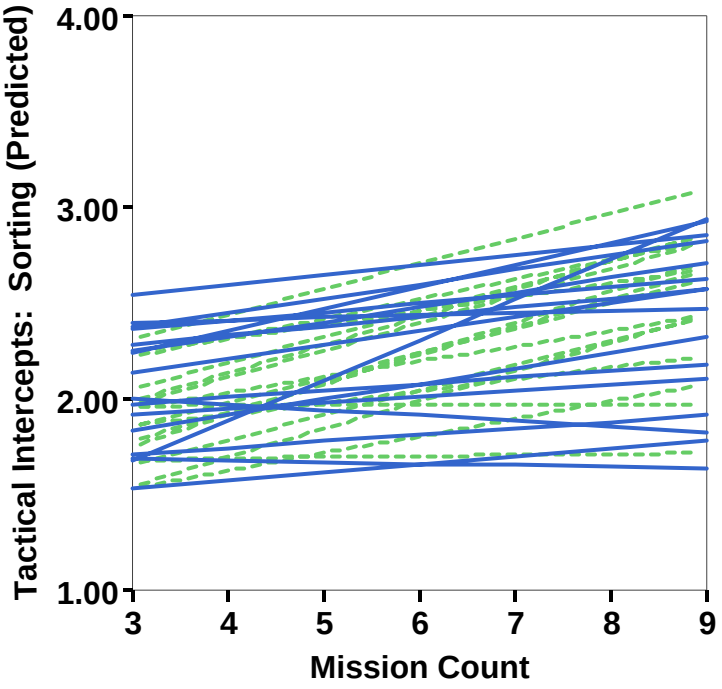

Gameplan: Tactics (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

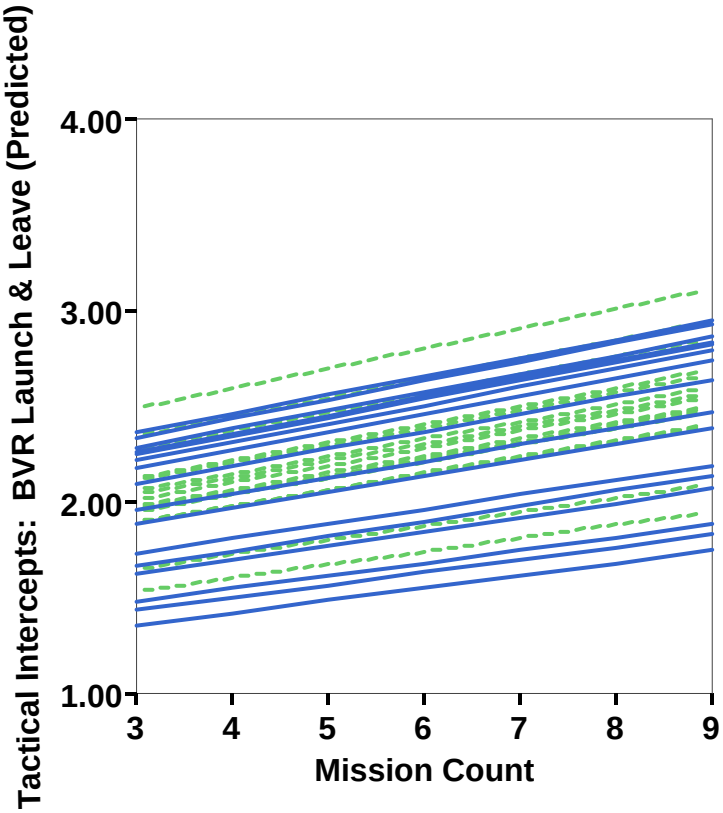

Gameplan: Execution (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

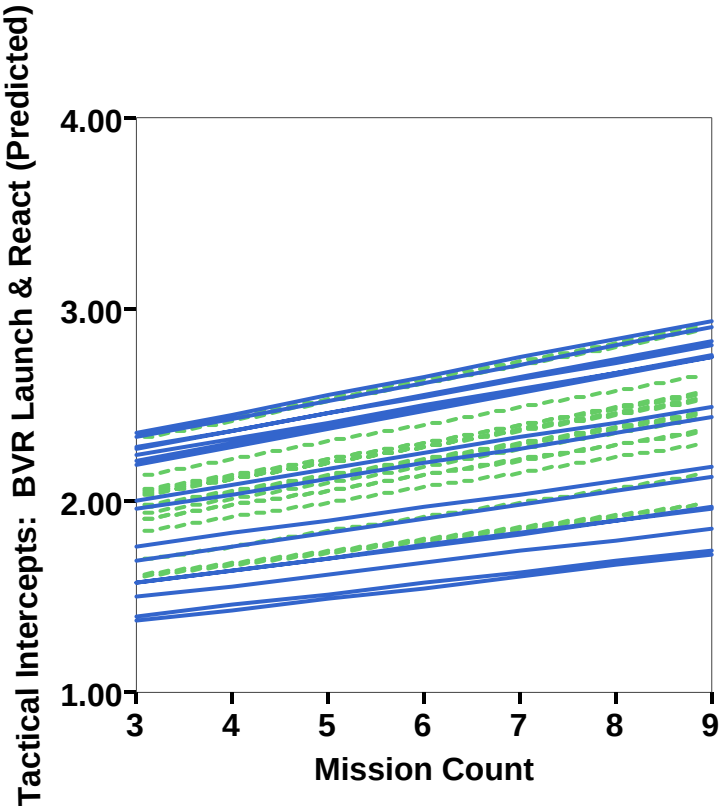

Gameplan: Adjusting Plan On-the-Fly (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count
Tactical Intercepts: Formation (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

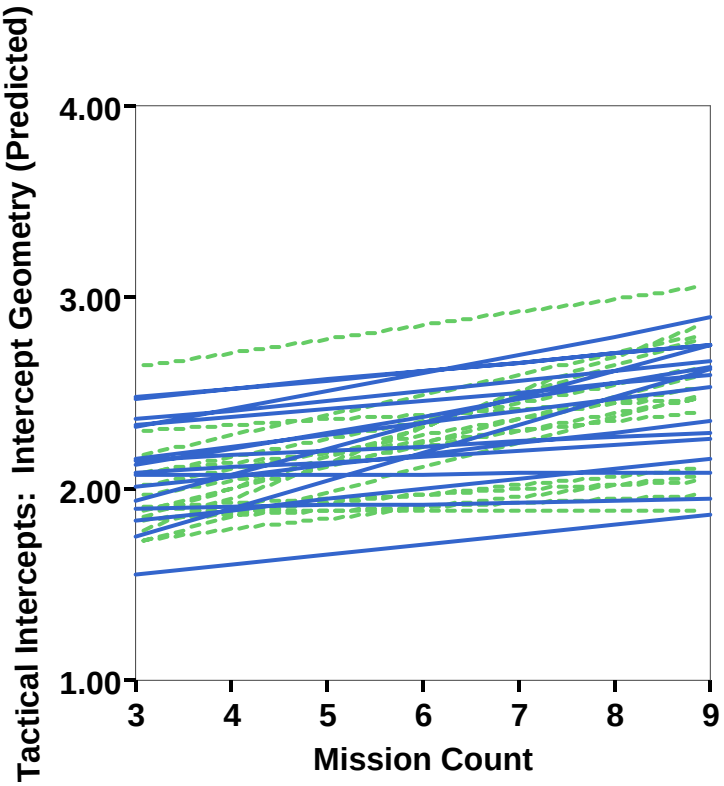

Tactical Intercepts: Detection - Commit (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

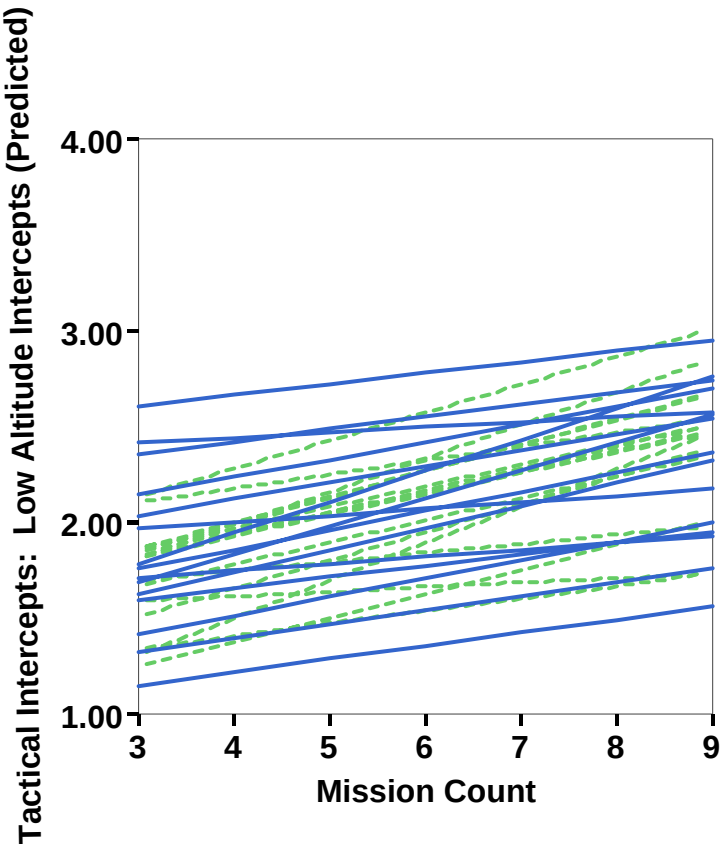

Tactical Intercepts: Targeting (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

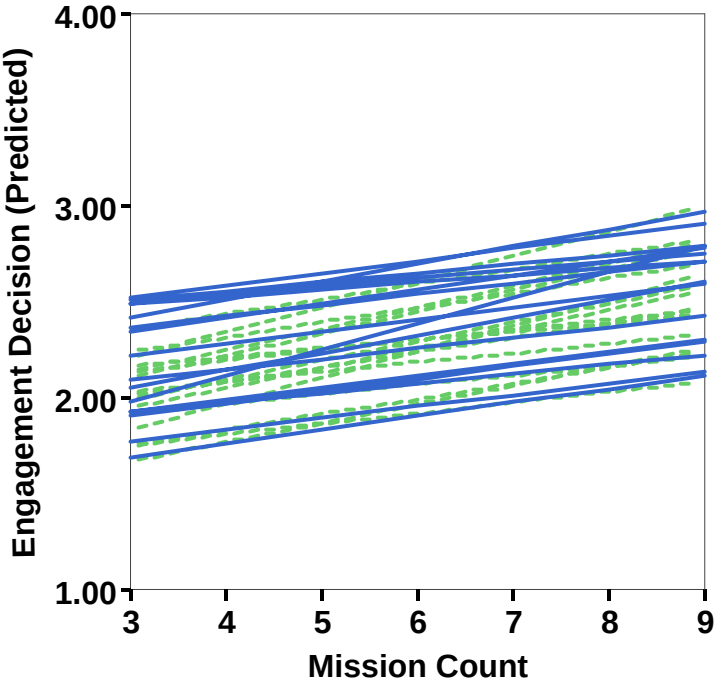

Tactical Intercepts: Sorting (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

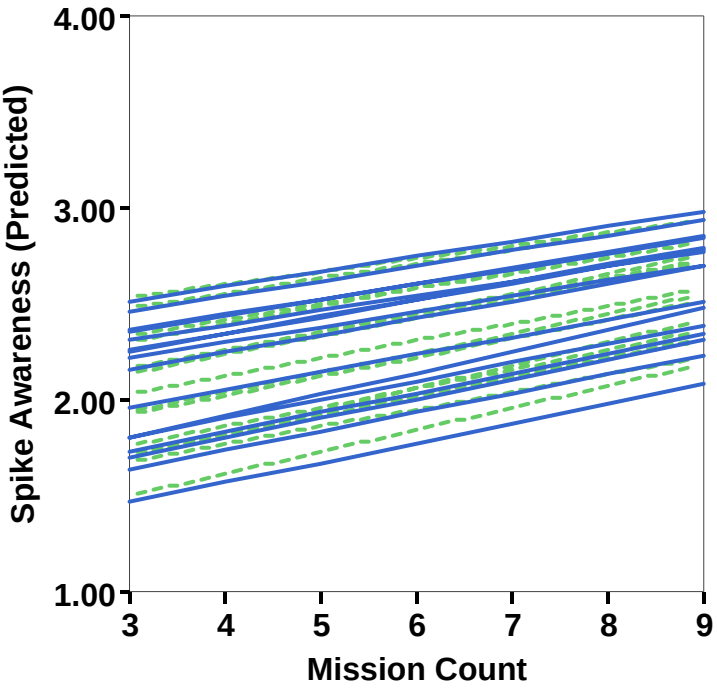

Tactical Intercepts: BVR Launch & Leave (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

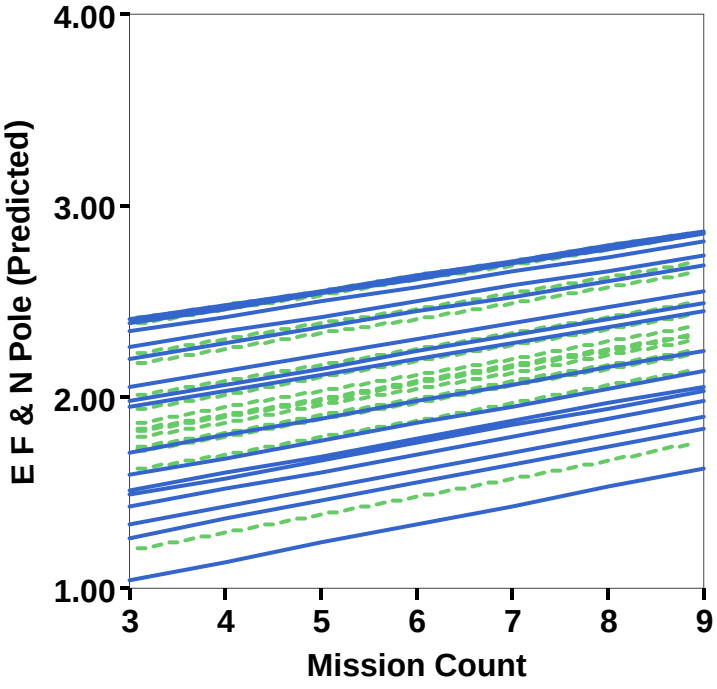

Tactical Intercepts: BVR Launch & React (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

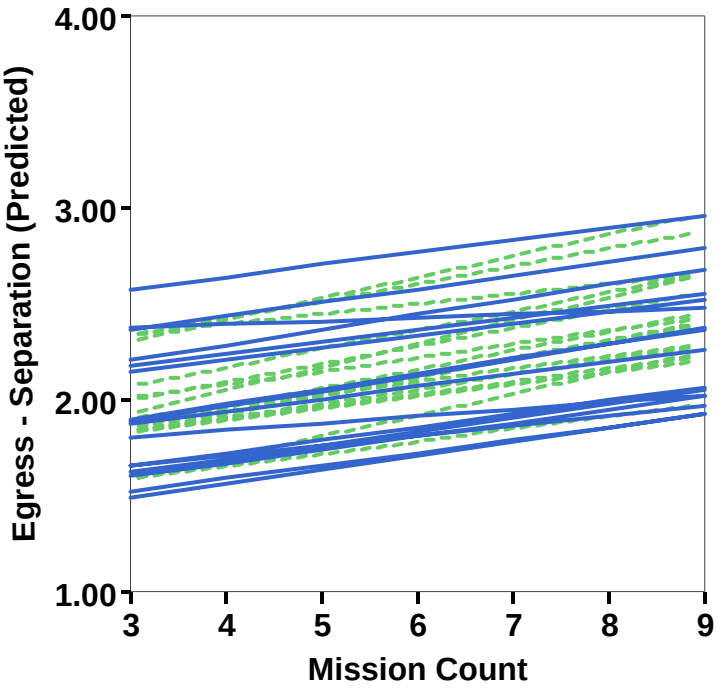

Tactical Intercepts: Intercept Geometry (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

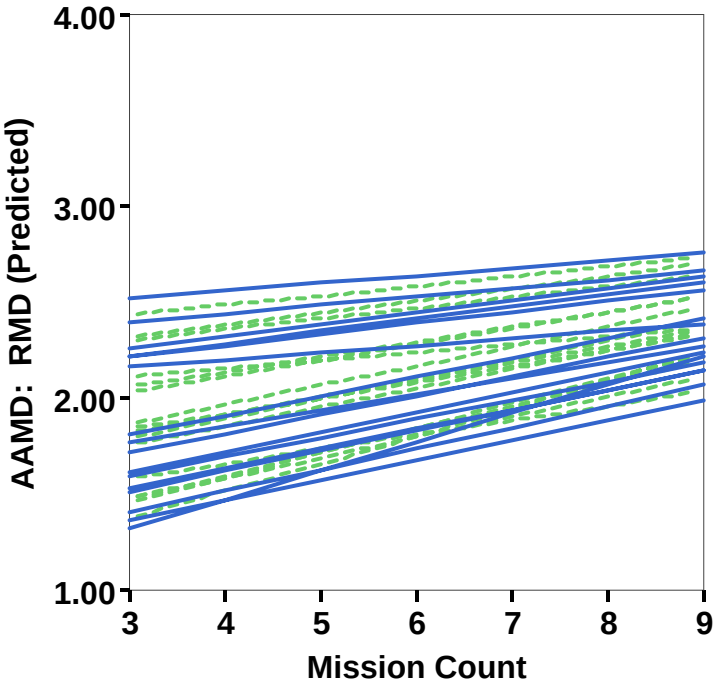

Tactical Intercepts: Low Altitude Intercepts (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

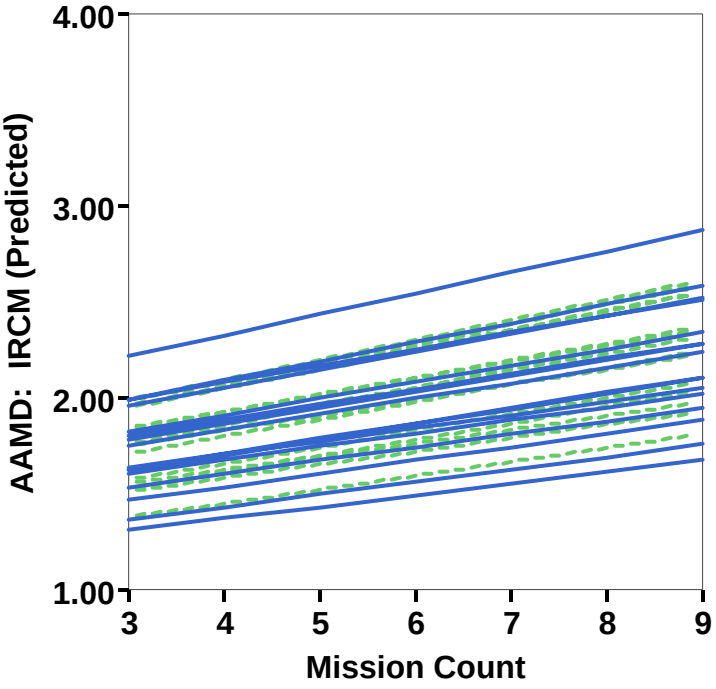

Engagement Decision (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

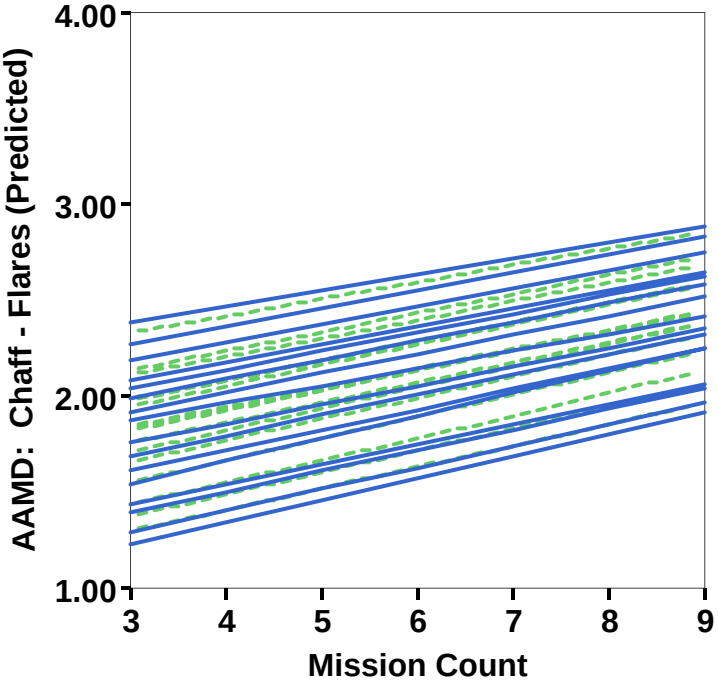

Spike Awareness (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

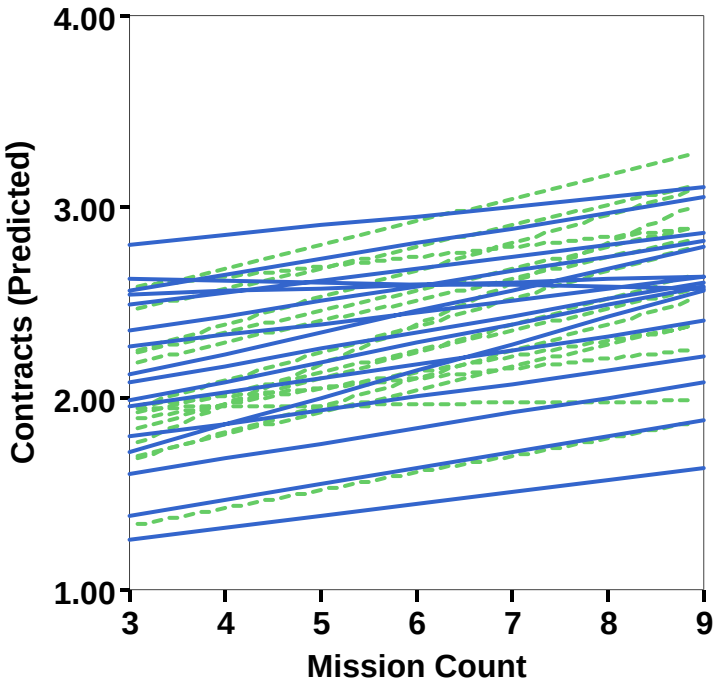

E F & N Pole (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

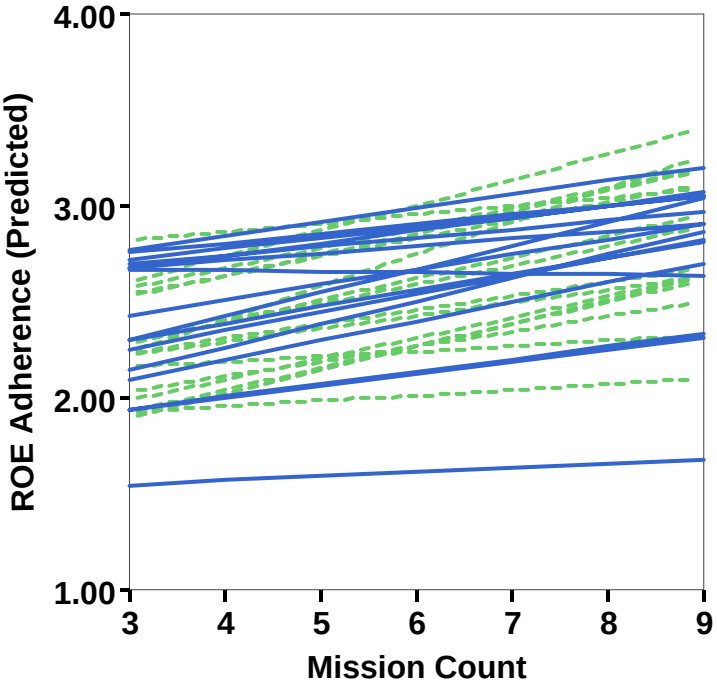

4.00
3.00
2.00
1.00
Egress - Separation (Predicted)
3
4
5
6
7
8
9
Mission Count

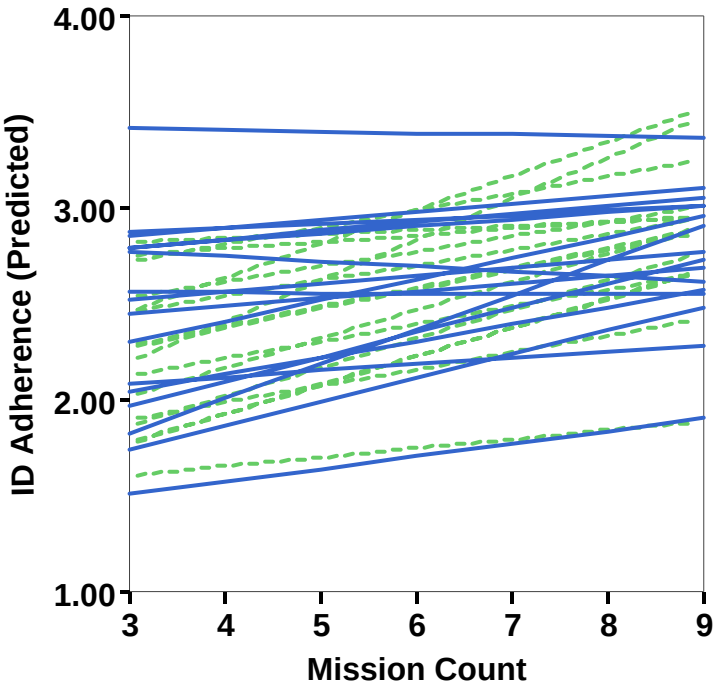

4.00
3.00
2.00
1.00
AAMD: RMD (Predicted)
3
4
5
6
7
8
9
Mission Count

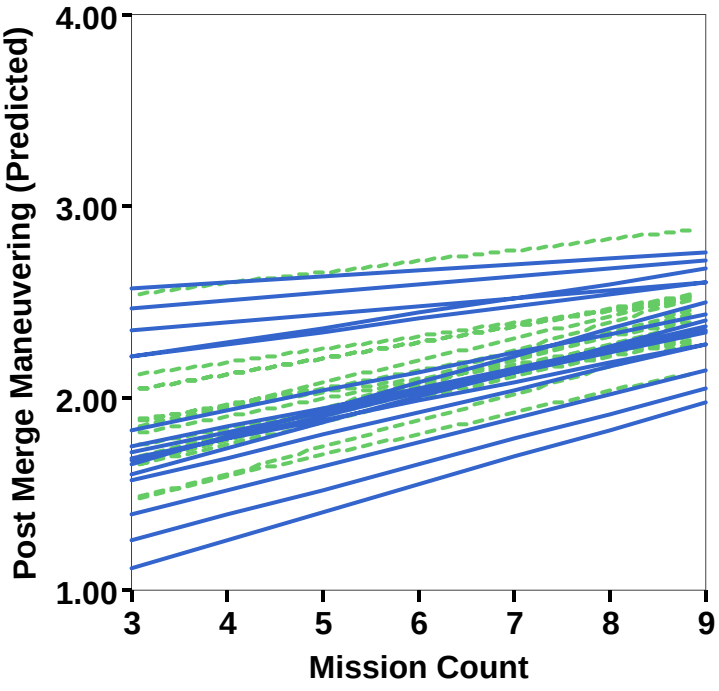

4.00
3.00
2.00
1.00
AAMD: IRCM (Predicted)
3
4
5
6
7
8
9
Mission Count

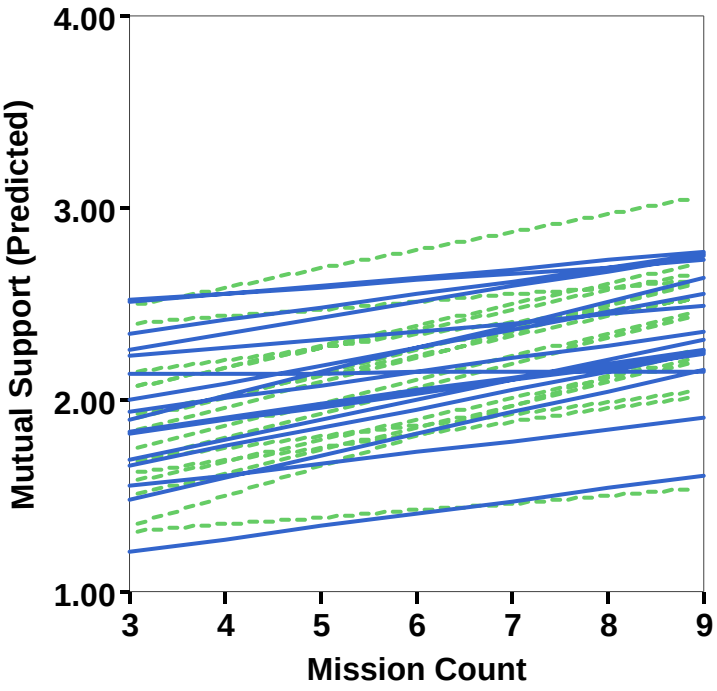

4.00
3.00
2.00
1.00
AAMD: Chaff - Flares (Predicted)
3
4
5
6
7
8
9
Mission Count

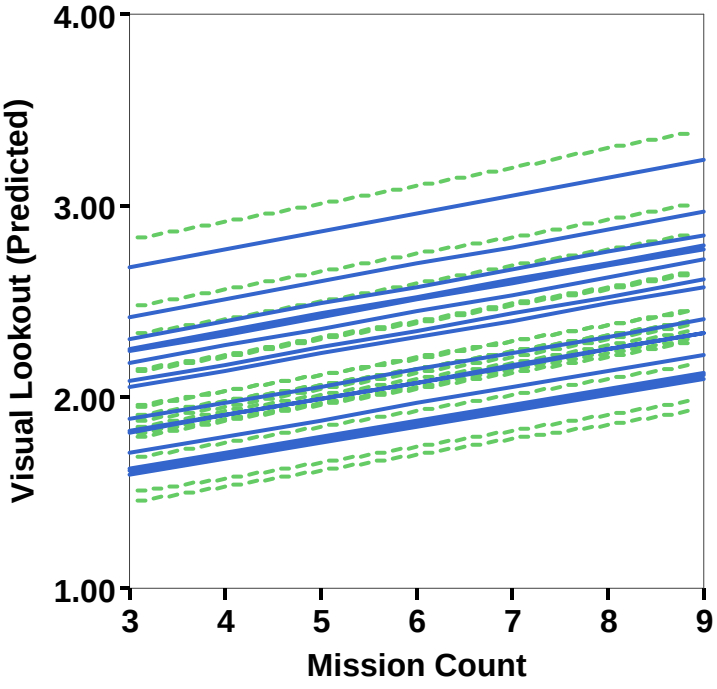

4.00
3.00
2.00
1.00
Contracts (Predicted)
3
4
5
6
7
8
9
Mission Count

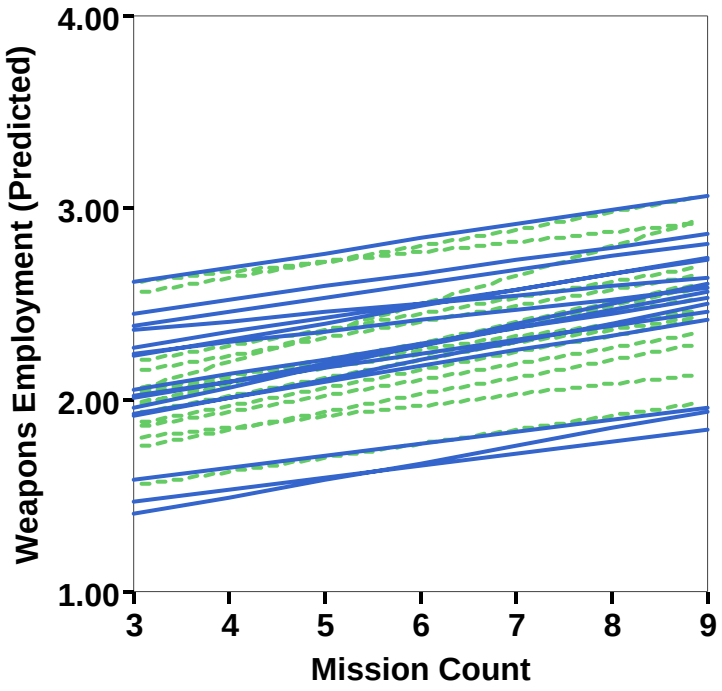

4.00
3.00
2.00
1.00
ROE Adherence (Predicted)
3
4
5
6
7
8
9
Mission Count

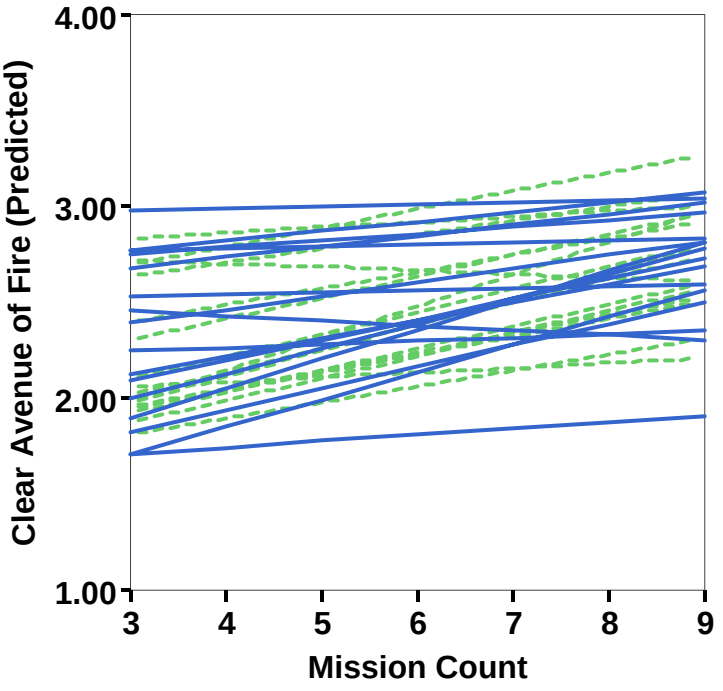

4.00
3.00
2.00
1.00
ID Adherence (Predicted)
3
4
5
6
7
8
9
Mission Count

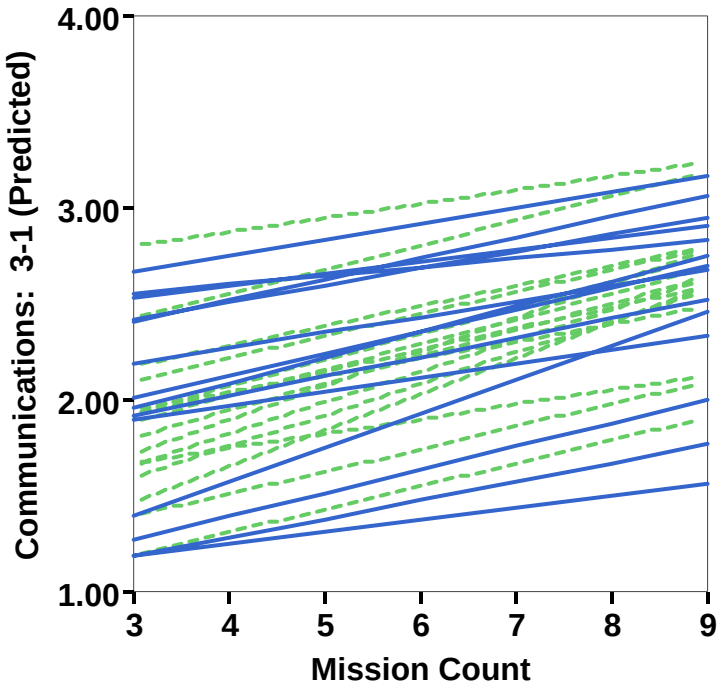

4.00
3.00
2.00
1.00
Post Merge Maneuvering (Predicted)
3
4
5
6
7
8
9
Mission Count

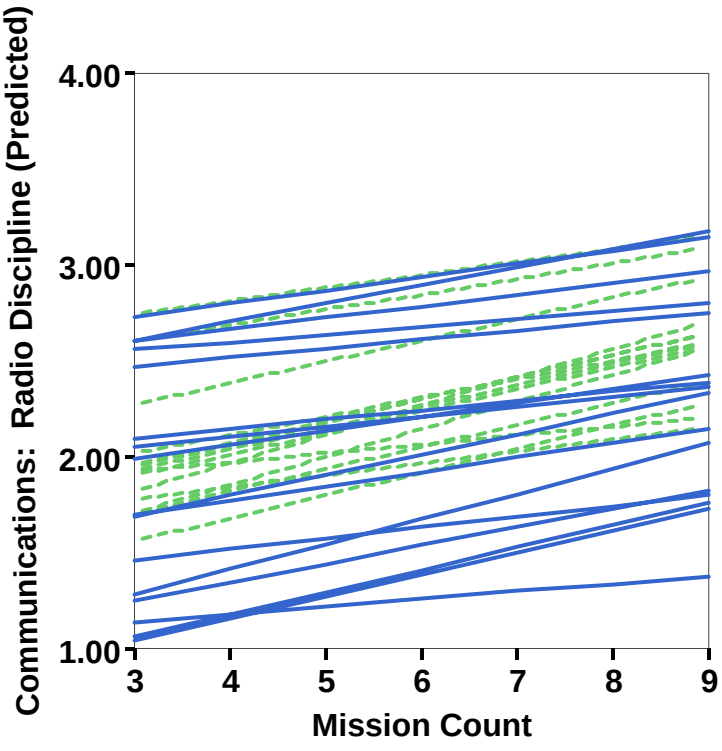

4.00
3.00
2.00
1.00
Mutual Support (Predicted)
3
4
5
6
7
8
9
Mission Count

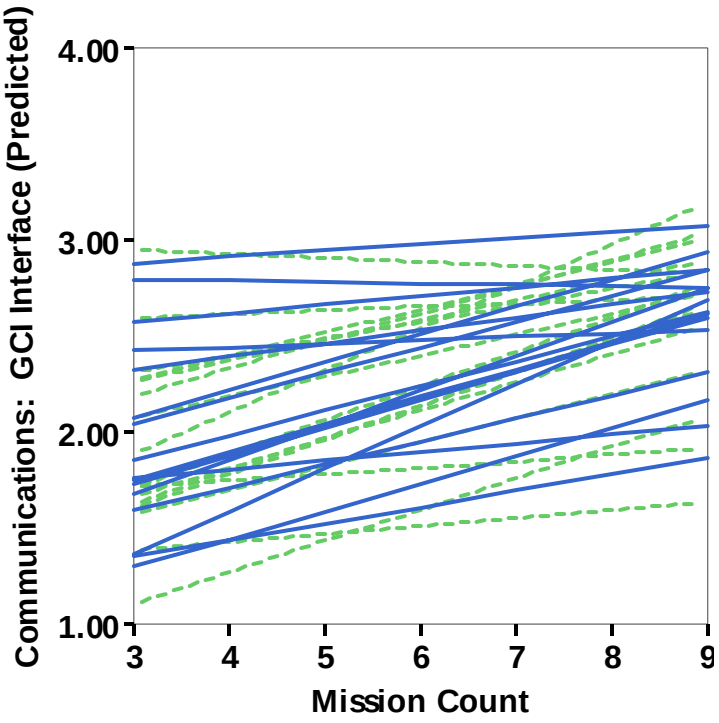

4.00
3.00
2.00
1.00
Visual Lookout (Predicted)
3
4
5
6
7
8
9
Mission Count

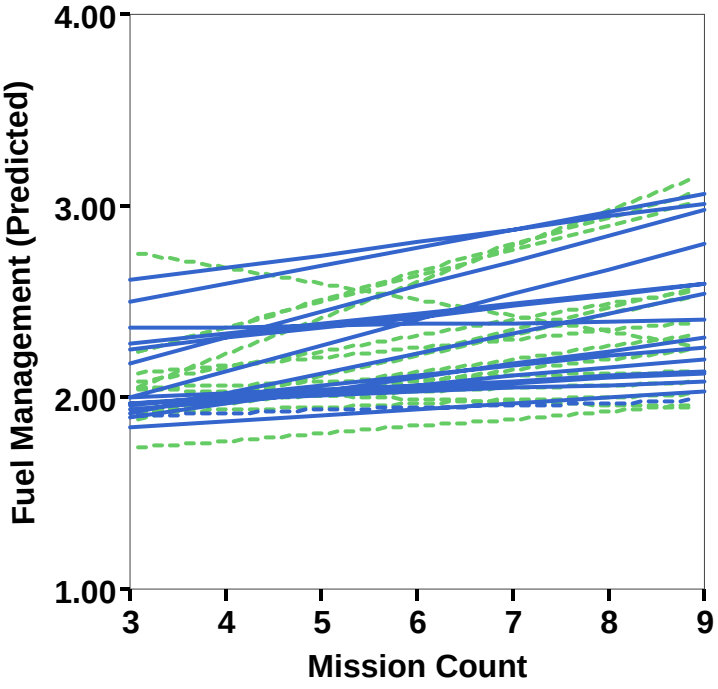

Weapons Employment (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

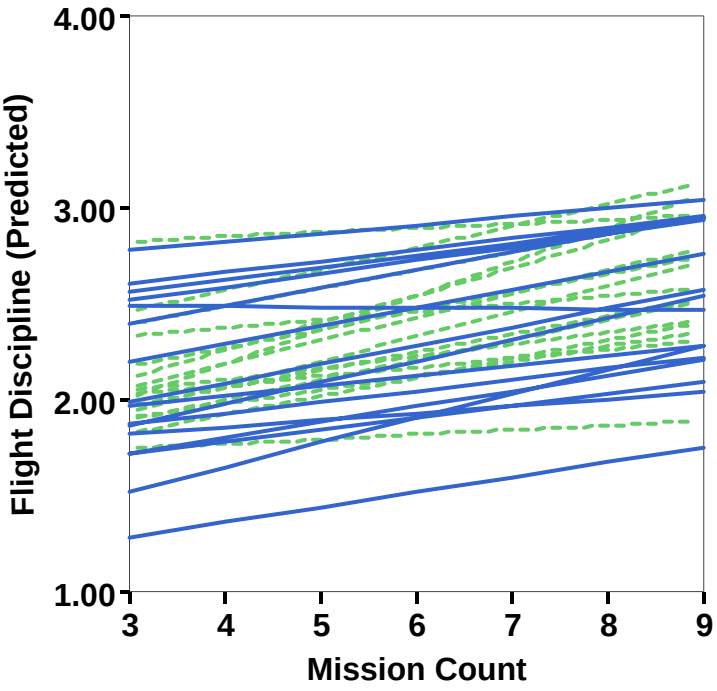

Clear Avenue of Fire (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

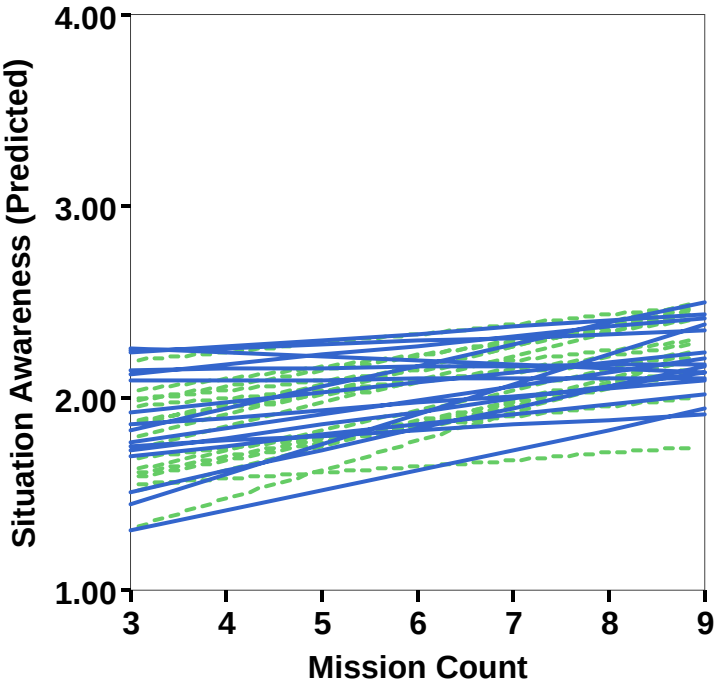

Communications: 3-1 (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

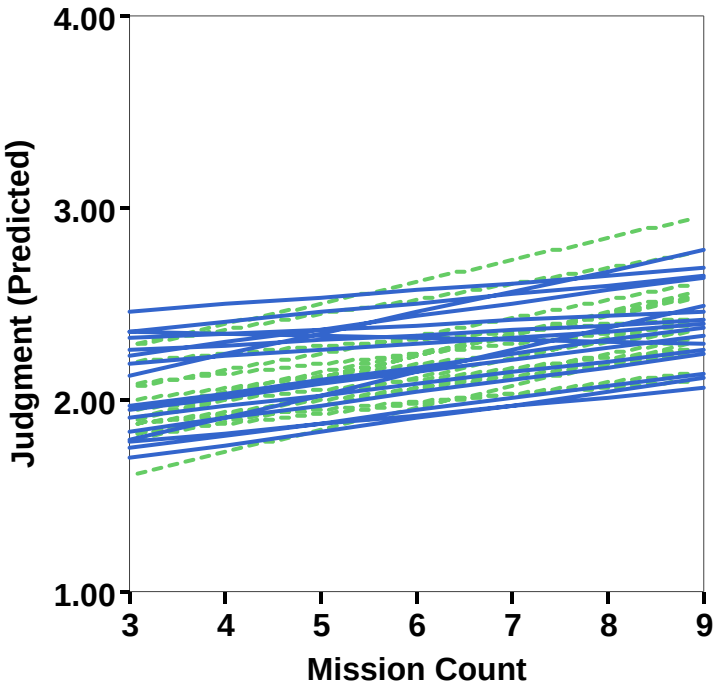

Communications: Radio Discipline (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

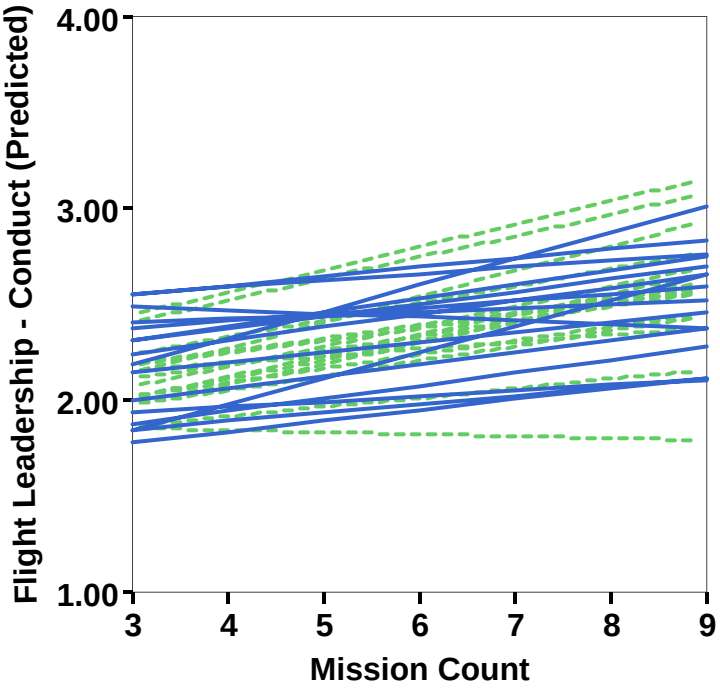

Communications: GCI Interface (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

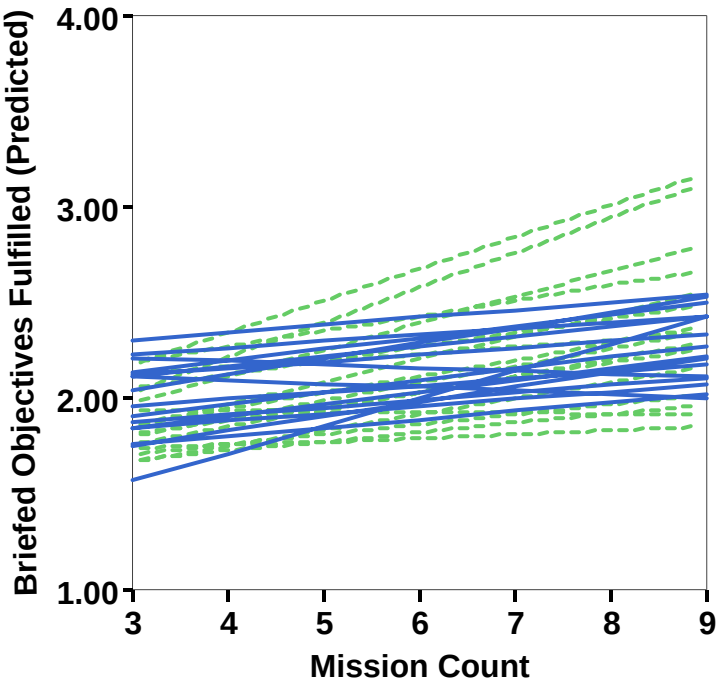

Fuel Management (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

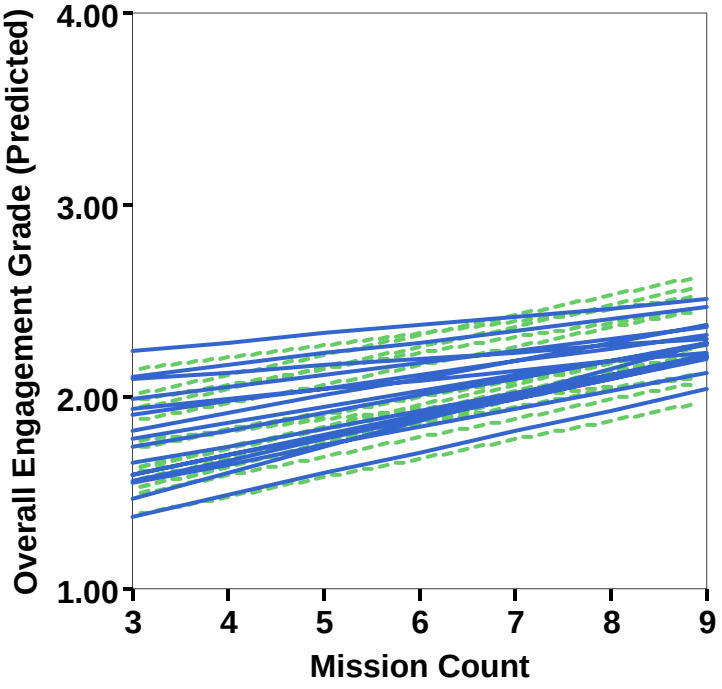

4.00
3.00
2.00
1.00
Flight Discipline (Predicted)
3
4
5
6
7
8
9
Mission Count


4.00
3.00
2.00
1.00
Situation Awareness (Predicted)
3
4
5
6
7
8
9
Mission Count


4.00
3.00
2.00
1.00
Judgment (Predicted)
3
4
5
6
7
8
9
Mission Count


4.00
3.00
2.00
1.00
Flight Leadership - Conduct (Predicted)
3
4
5
6
7
8
9
Mission Count


Briefed Objectives Fulfilled (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count


Overall Engagement Grade (Predicted)
4.00
3.00
2.00
1.00
3
4
5
6
7
8
9
Mission Count

Appendix E
Tables

Table 7a. **Correlations among graded performance indicators.**

Indicator	1	2	3	4	5	6	7	8	9	10
Radar Mechanics:										
1. El Strobe Control	1.00	0.81	0.74	0.69	0.72	0.59	0.58	0.73	0.76	0.63
2. Range Control	0.81	1.00	0.88	0.74	0.80	0.59	0.55	0.66	0.80	0.72
3. Azimuth Control	0.74	0.88	1.00	0.78	0.81	0.61	0.52	0.65	0.80	0.73
4. Utilizing Correct Mode	0.69	0.74	0.78	1.00	0.76	0.57	0.54	0.65	0.71	0.66
Gameplan:										
5. Tactics	0.72	0.80	0.81	0.76	1.00	0.69	0.65	0.68	0.83	0.76
6. Execution	0.59	0.59	0.61	0.57	0.69	1.00	0.74	0.62	0.72	0.75
7. Adjusting Plan On-the-Fly	0.58	0.55	0.52	0.54	0.65	0.74	1.00	0.69	0.65	0.59
Tactical Intercepts:										
8. Formation	0.73	0.66	0.65	0.65	0.68	0.62	0.69	1.00	0.75	0.59
9. Detection – Commit	0.76	0.80	0.80	0.71	0.83	0.72	0.65	0.75	1.00	0.77
10. Targeting	0.63	0.72	0.73	0.66	0.76	0.75	0.59	0.59	0.77	1.00
11. Sorting	0.68	0.67	0.69	0.67	0.71	0.69	0.61	0.62	0.73	0.83
12. BVR Launch and Leave	0.61	0.62	0.59	0.63	0.67	0.70	0.65	0.59	0.66	0.70
13. BVR Launch and React	0.66	0.68	0.67	0.61	0.68	0.68	0.60	0.58	0.70	0.70
14. Intercept Geometry	0.63	0.60	0.62	0.69	0.61	0.61	0.64	0.62	0.67	0.56
15. Low Altitude Intercepts	0.54	0.57	0.58	0.62	0.60	0.60	0.59	0.62	0.62	0.56
AAMD:										
16. RMD	0.47	0.57	0.60	0.56	0.64	0.59	0.60	0.48	0.60	0.57
17. IRCM	0.43	0.49	0.49	0.43	0.54	0.51	0.52	0.45	0.52	0.48
18. Chaff – Flares	0.47	0.53	0.50	0.43	0.58	0.52	0.53	0.43	0.54	0.50
Communications:										
19. 3-1 Communication	0.61	0.59	0.54	0.57	0.72	0.57	0.68	0.64	0.65	0.58
20. Radio Discipline	0.57	0.59	0.54	0.59	0.69	0.56	0.68	0.61	0.64	0.59
21. GCI Interface	0.54	0.56	0.57	0.59	0.62	0.59	0.65	0.58	0.66	0.58
Additional Indicators										
22. Engagement Decision	0.66	0.66	0.65	0.63	0.72	0.75	0.71	0.68	0.77	0.69
23. Spike Awareness	0.54	0.62	0.62	0.59	0.68	0.64	0.57	0.56	0.68	0.61
24. E F & N Pole	0.53	0.54	0.58	0.54	0.67	0.64	0.65	0.60	0.63	0.59
25. Egress – Separation	0.49	0.52	0.52	0.50	0.65	0.65	0.70	0.56	0.60	0.55
26. Contracts	0.73	0.67	0.69	0.69	0.75	0.69	0.72	0.75	0.75	0.66
27. ROE Adherence	0.60	0.56	0.60	0.66	0.69	0.62	0.66	0.73	0.71	0.51
28. ID Adherence	0.52	0.58	0.61	0.63	0.67	0.52	0.48	0.59	0.62	0.54
29. Post Merge Maneuvering	0.48	0.51	0.49	0.55	0.62	0.57	0.59	0.59	0.59	0.48
30. Mutual Support	0.61	0.60	0.58	0.58	0.65	0.67	0.72	0.77	0.70	0.58
31. Visual Lookout	0.67	0.61	0.60	0.63	0.69	0.58	0.66	0.75	0.66	0.53
32. Weapons Employment	0.57	0.60	0.62	0.67	0.65	0.56	0.58	0.63	0.68	0.57
33. Clear Avenue of Fire	0.48	0.55	0.60	0.64	0.62	0.56	0.48	0.55	0.63	0.54
34. Fuel Management	0.52	0.54	0.49	0.48	0.52	0.56	0.52	0.52	0.52	0.55
35. Flight Discipline	0.61	0.65	0.59	0.64	0.71	0.67	0.69	0.70	0.69	0.66
36. Situation Awareness	0.50	0.52	0.49	0.54	0.63	0.78	0.78	0.59	0.64	0.63
37. Judgment	0.51	0.54	0.56	0.56	0.62	0.75	0.76	0.62	0.65	0.62
38. Flight Leadership – Conduct	0.55	0.57	0.56	0.57	0.65	0.68	0.74	0.62	0.68	0.59
39. Briefed Objectives Fulfilled	0.37	0.40	0.42	0.44	0.46	0.63	0.66	0.50	0.53	0.50
40. Overall Engagement Grade	0.51	0.52	0.54	0.61	0.60	0.76	0.72	0.59	0.63	0.63

Table 7b. **Correlations among graded performance indicators.**

Indicator	11	12	13	14	15	16	17	18	19	20
Radar Mechanics:										
1. El Strobe Control	0.68	0.61	0.66	0.63	0.54	0.47	0.43	0.47	0.61	0.57
2. Range Control	0.67	0.62	0.68	0.60	0.57	0.57	0.49	0.53	0.59	0.59
3. Azimuth Control	0.69	0.59	0.67	0.62	0.58	0.60	0.49	0.50	0.54	0.54
4. Utilizing Correct Mode	0.67	0.63	0.61	0.69	0.62	0.56	0.43	0.43	0.57	0.59
Gameplan:										
5. Tactics	0.71	0.67	0.68	0.61	0.60	0.64	0.54	0.58	0.72	0.69
6. Execution	0.69	0.70	0.68	0.61	0.60	0.59	0.51	0.52	0.57	0.56
7. Adjusting Plan On-the-Fly	0.61	0.65	0.60	0.64	0.59	0.60	0.52	0.53	0.68	0.68
Tactical Intercepts:										
8. Formation	0.62	0.59	0.58	0.62	0.62	0.48	0.45	0.43	0.64	0.61
9. Detection – Commit	0.73	0.66	0.70	0.67	0.62	0.60	0.52	0.54	0.65	0.64
10. Targeting	0.83	0.70	0.70	0.56	0.56	0.57	0.48	0.50	0.58	0.59
11. Sorting	1.00	0.69	0.72	0.66	0.60	0.55	0.46	0.44	0.62	0.57
12. BVR Launch and Leave	0.69	1.00	0.86	0.60	0.63	0.69	0.60	0.64	0.64	0.64
13. BVR Launch and React	0.72	0.86	1.00	0.64	0.62	0.74	0.63	0.67	0.62	0.60
14. Intercept Geometry	0.66	0.60	0.64	1.00	0.64	0.51	0.43	0.38	0.55	0.58
15. Low Altitude Intercepts	0.60	0.63	0.62	0.64	1.00	0.59	0.59	0.50	0.61	0.60
AAMD:										
16. RMD	0.55	0.69	0.74	0.51	0.59	1.00	0.74	0.78	0.58	0.61
17. IRCM	0.46	0.60	0.63	0.43	0.59	0.74	1.00	0.78	0.51	0.49
18. Chaff – Flares	0.44	0.64	0.67	0.38	0.50	0.78	0.78	1.00	0.58	0.55
Communications:										
19. 3-1 Communication	0.62	0.64	0.62	0.55	0.61	0.58	0.51	0.58	1.00	0.85
20. Radio Discipline	0.57	0.64	0.60	0.58	0.60	0.61	0.49	0.55	0.85	1.00
21. GCI Interface	0.57	0.59	0.59	0.62	0.58	0.57	0.47	0.51	0.68	0.70
Additional Indicators										
22. Engagement Decision	0.66	0.69	0.68	0.66	0.61	0.58	0.51	0.51	0.63	0.61
23. Spike Awareness	0.63	0.72	0.76	0.62	0.60	0.77	0.64	0.72	0.65	0.60
24. E F & N Pole	0.61	0.73	0.74	0.57	0.64	0.72	0.59	0.61	0.71	0.70
25. Egress – Separation	0.57	0.73	0.73	0.55	0.66	0.76	0.64	0.63	0.68	0.67
26. Contracts	0.71	0.71	0.67	0.69	0.68	0.62	0.61	0.60	0.73	0.70
27. ROE Adherence	0.50	0.58	0.52	0.61	0.59	0.53	0.54	0.53	0.65	0.62
28. ID Adherence	0.44	0.58	0.51	0.48	0.49	0.54	0.51	0.56	0.60	0.64
29. Post Merge Maneuvering	0.52	0.56	0.57	0.56	0.67	0.62	0.61	0.52	0.62	0.62
30. Mutual Support	0.59	0.64	0.63	0.64	0.68	0.61	0.58	0.57	0.67	0.69
31. Visual Lookout	0.63	0.61	0.63	0.70	0.60	0.57	0.54	0.50	0.69	0.62
32. Weapons Employment	0.64	0.63	0.66	0.59	0.57	0.57	0.49	0.48	0.63	0.62
33. Clear Avenue of Fire	0.50	0.57	0.52	0.55	0.55	0.57	0.51	0.48	0.58	0.62
34. Fuel Management	0.53	0.60	0.57	0.49	0.53	0.57	0.45	0.55	0.58	0.57
35. Flight Discipline	0.65	0.72	0.68	0.59	0.65	0.66	0.59	0.60	0.73	0.68
36. Situation Awareness	0.56	0.65	0.60	0.59	0.56	0.62	0.52	0.54	0.59	0.67
37. Judgment	0.57	0.60	0.57	0.62	0.60	0.57	0.47	0.49	0.57	0.60
38. Flight Leadership – Conduct	0.51	0.56	0.52	0.61	0.56	0.56	0.44	0.47	0.55	0.64
39. Briefed Objectives Fulfilled	0.40	0.44	0.40	0.51	0.45	0.48	0.35	0.36	0.42	0.54
40. Overall Engagement Grade	0.57	0.63	0.59	0.63	0.57	0.62	0.51	0.44	0.52	0.57

Table 7c. **Correlations among graded performance indicators.**

Indicator	21	22	23	24	25	26	27	28	29	30
Radar Mechanics:										
1. El Strobe Control	0.54	0.66	0.54	0.53	0.49	0.73	0.60	0.52	0.48	0.61
2. Range Control	0.56	0.66	0.62	0.54	0.52	0.67	0.56	0.58	0.51	0.60
3. Azimuth Control	0.57	0.65	0.62	0.58	0.52	0.69	0.60	0.61	0.49	0.58
4. Utilizing Correct Mode	0.59	0.63	0.59	0.54	0.50	0.69	0.66	0.63	0.55	0.58
Gameplan:										
5. Tactics	0.62	0.72	0.68	0.67	0.65	0.75	0.69	0.67	0.62	0.65
6. Execution	0.59	0.75	0.64	0.64	0.65	0.69	0.62	0.52	0.57	0.67
7. Adjusting Plan On-the-Fly	0.65	0.71	0.57	0.65	0.70	0.72	0.66	0.48	0.59	0.72
Tactical Intercepts:										
8. Formation	0.58	0.68	0.56	0.60	0.56	0.75	0.73	0.59	0.59	0.77
9. Detection – Commit	0.66	0.77	0.68	0.63	0.60	0.75	0.71	0.62	0.59	0.70
10. Targeting	0.58	0.69	0.61	0.59	0.55	0.66	0.51	0.54	0.48	0.58
11. Sorting	0.57	0.66	0.63	0.61	0.57	0.71	0.50	0.44	0.52	0.59
12. BVR Launch and Leave	0.59	0.69	0.72	0.73	0.73	0.71	0.58	0.58	0.56	0.64
13. BVR Launch and React	0.59	0.68	0.76	0.74	0.73	0.67	0.52	0.51	0.57	0.63
14. Intercept Geometry	0.62	0.66	0.62	0.57	0.55	0.69	0.61	0.48	0.56	0.64
15. Low Altitude Intercepts	0.58	0.61	0.60	0.64	0.66	0.68	0.59	0.49	0.67	0.68
AAMD:										
16. RMD	0.57	0.58	0.77	0.72	0.76	0.62	0.53	0.54	0.62	0.61
17. IRCM	0.47	0.51	0.64	0.59	0.64	0.61	0.54	0.51	0.61	0.58
18. Chaff – Flares	0.51	0.51	0.72	0.61	0.63	0.60	0.53	0.56	0.52	0.57
Communications:										
19. 3-1 Communication	0.68	0.63	0.65	0.71	0.68	0.73	0.65	0.60	0.62	0.67
20. Radio Discipline	0.70	0.61	0.60	0.70	0.67	0.70	0.62	0.64	0.62	0.69
21. GCI Interface	1.00	0.68	0.61	0.65	0.60	0.68	0.63	0.55	0.54	0.68
Additional Indicators										
22. Engagement Decision	0.68	1.00	0.65	0.60	0.62	0.76	0.70	0.57	0.55	0.67
23. Spike Awareness	0.61	0.65	1.00	0.75	0.72	0.63	0.65	0.62	0.60	0.63
24. E F & N Pole	0.65	0.60	0.75	1.00	0.82	0.71	0.61	0.57	0.65	0.62
25. Egress – Separation	0.60	0.62	0.72	0.82	1.00	0.68	0.61	0.57	0.74	0.63
26. Contracts	0.68	0.76	0.63	0.71	0.68	1.00	0.79	0.62	0.66	0.78
27. ROE Adherence	0.63	0.70	0.65	0.61	0.61	0.79	1.00	0.79	0.64	0.71
28. ID Adherence	0.55	0.57	0.62	0.57	0.57	0.62	0.79	1.00	0.61	0.61
29. Post Merge Maneuvering	0.54	0.55	0.60	0.65	0.74	0.66	0.64	0.61	1.00	0.73
30. Mutual Support	0.68	0.67	0.63	0.62	0.63	0.78	0.71	0.61	0.73	1.00
31. Visual Lookout	0.62	0.65	0.69	0.65	0.63	0.71	0.69	0.54	0.66	0.74
32. Weapons Employment	0.67	0.62	0.61	0.64	0.61	0.69	0.69	0.58	0.61	0.57
33. Clear Avenue of Fire	0.59	0.58	0.61	0.59	0.59	0.62	0.65	0.69	0.64	0.63
34. Fuel Management	0.48	0.56	0.60	0.59	0.61	0.58	0.56	0.52	0.56	0.53
35. Flight Discipline	0.66	0.69	0.72	0.69	0.69	0.70	0.71	0.64	0.67	0.73
36. Situation Awareness	0.67	0.71	0.64	0.61	0.65	0.64	0.65	0.57	0.62	0.73
37. Judgment	0.65	0.72	0.61	0.56	0.62	0.63	0.64	0.53	0.55	0.66
38. Flight Leadership – Conduct	0.67	0.73	0.54	0.55	0.58	0.64	0.64	0.54	0.52	0.67
39. Briefed Objectives Fulfilled	0.61	0.60	0.50	0.48	0.54	0.53	0.58	0.47	0.48	0.57
40. Overall Engagement Grade	0.68	0.73	0.62	0.60	0.63	0.66	0.66	0.54	0.62	0.66

Table 7d. **Correlations among graded performance indicators.**

Indicator	31	32	33	34	35	36	37	38	39	40
Radar Mechanics:										
1. El Strobe Control	0.67	0.57	0.48	0.52	0.61	0.50	0.51	0.55	0.37	0.51
2. Range Control	0.61	0.60	0.55	0.54	0.65	0.52	0.54	0.57	0.40	0.52
3. Azimuth Control	0.60	0.62	0.60	0.49	0.59	0.49	0.56	0.56	0.42	0.54
4. Utilizing Correct Mode	0.63	0.67	0.64	0.48	0.64	0.54	0.56	0.57	0.44	0.61
Gameplan:										
5. Tactics	0.69	0.65	0.62	0.52	0.71	0.63	0.62	0.65	0.46	0.60
6. Execution	0.58	0.56	0.56	0.56	0.67	0.78	0.75	0.68	0.63	0.76
7. Adjusting Plan On-the-Fly	0.66	0.58	0.48	0.52	0.69	0.78	0.76	0.74	0.66	0.72
Tactical Intercepts:										
8. Formation	0.75	0.63	0.55	0.52	0.70	0.59	0.62	0.62	0.50	0.59
9. Detection – Commit	0.66	0.68	0.63	0.52	0.69	0.64	0.65	0.68	0.53	0.63
10. Targeting	0.53	0.57	0.54	0.55	0.66	0.63	0.62	0.59	0.50	0.63
11. Sorting	0.63	0.64	0.50	0.53	0.65	0.56	0.57	0.51	0.40	0.57
12. BVR Launch and Leave	0.61	0.63	0.57	0.60	0.72	0.65	0.60	0.56	0.44	0.63
13. BVR Launch and React	0.63	0.66	0.52	0.57	0.68	0.60	0.57	0.52	0.40	0.59
14. Intercept Geometry	0.70	0.59	0.55	0.49	0.59	0.59	0.62	0.61	0.51	0.63
15. Low Altitude Intercepts	0.60	0.57	0.55	0.53	0.65	0.56	0.60	0.56	0.45	0.57
AAMD:										
16. RMD	0.57	0.57	0.57	0.57	0.66	0.62	0.57	0.56	0.48	0.62
17. IRCM	0.54	0.49	0.51	0.45	0.59	0.52	0.47	0.44	0.35	0.51
18. Chaff – Flares	0.50	0.48	0.48	0.55	0.60	0.54	0.49	0.47	0.36	0.44
Communications:										
19. 3-1 Communication	0.69	0.63	0.58	0.58	0.73	0.59	0.57	0.55	0.42	0.52
20. Radio Discipline	0.62	0.62	0.62	0.57	0.68	0.67	0.60	0.64	0.54	0.57
21. GCI Interface	0.62	0.67	0.59	0.48	0.66	0.67	0.65	0.67	0.61	0.68
Additional Indicators										
22. Engagement Decision	0.65	0.62	0.58	0.56	0.69	0.71	0.72	0.73	0.60	0.73
23. Spike Awareness	0.69	0.61	0.61	0.60	0.72	0.64	0.61	0.54	0.50	0.62
24. E F & N Pole	0.65	0.64	0.59	0.59	0.69	0.61	0.56	0.55	0.48	0.60
25. Egress – Separation	0.63	0.61	0.59	0.61	0.69	0.65	0.62	0.58	0.54	0.63
26. Contracts	0.71	0.69	0.62	0.58	0.70	0.64	0.63	0.64	0.53	0.66
27. ROE Adherence	0.69	0.69	0.65	0.56	0.71	0.65	0.64	0.64	0.58	0.66
28. ID Adherence	0.54	0.58	0.69	0.52	0.64	0.57	0.53	0.54	0.47	0.54
29. Post Merge Maneuvering	0.66	0.61	0.64	0.56	0.67	0.62	0.55	0.52	0.48	0.62
30. Mutual Support	0.74	0.57	0.63	0.53	0.73	0.73	0.66	0.67	0.57	0.66
31. Visual Lookout	1.00	0.63	0.58	0.55	0.76	0.62	0.58	0.58	0.48	0.62
32. Weapons Employment	0.63	1.00	0.64	0.54	0.67	0.57	0.57	0.55	0.49	0.60
33. Clear Avenue of Fire	0.58	0.64	1.00	0.46	0.60	0.55	0.57	0.54	0.58	0.64
34. Fuel Management	0.55	0.54	0.46	1.00	0.67	0.48	0.56	0.51	0.46	0.51
35. Flight Discipline	0.76	0.67	0.60	0.67	1.00	0.69	0.70	0.65	0.54	0.66
36. Situation Awareness	0.62	0.57	0.55	0.48	0.69	1.00	0.83	0.78	0.72	0.79
37. Judgment	0.58	0.57	0.57	0.56	0.70	0.83	1.00	0.85	0.76	0.77
38. Flight Leadership – Conduct	0.58	0.55	0.54	0.51	0.65	0.78	0.85	1.00	0.77	0.75
39. Briefed Objectives Fulfilled	0.48	0.49	0.58	0.46	0.54	0.72	0.76	0.77	1.00	0.86
40. Overall Engagement Grade	0.62	0.60	0.64	0.51	0.66	0.79	0.77	0.75	0.86	1.00

Table 8. **Component loadings, communalities, eigenvalue, and percent of variance accounted for computed from a principal components analysis on all indicators of graded 4-ship team performance.**

Indicator	Component Loadings	Communalities
Radar Mechanics:		
El Strobe Control	.759	.576
Range Control	.787	.619
Azimuth Control	.784	.615
Utilizing Correct Mode	.781	.610
Gameplan:		
Tactics	.857	.734
Execution	.818	.669
Adjusting Plan On-the-Fly	.813	.662
Tactical Intercepts:		
Formation	.799	.638
Detection – Commit	.861	.742
Targeting	.790	.624
Sorting	.780	.608
BVR Launch and Leave	.823	.677
BVR Launch and React	.818	.669
Intercept Geometry	.765	.586
Low Altitude Intercepts	.763	.583
AAMD:		
RMD	.777	.603
IRCM	.684	.467
Chaff – Flares	.695	.483
Communications:		
3-1 Communication	.799	.638
Radio Discipline	.799	.638
GCI Interface	.784	.614
Additional Indicators		
Engagement Decision	.839	.703
Spike Awareness	.817	.668
E F & N Pole	.809	.654
Egress – Separation	.807	.651
Contracts	.873	.762
ROE Adherence	.812	.659
ID Adherence	.735	.540
Post Merge Maneuvering	.759	.576
Mutual Support	.835	.697
Visual Lookout	.807	.652
Weapons Employment	.778	.606
Clear Avenue of Fire	.743	.552
Fuel Management	.694	.482
Flight Discipline	.856	.733
Situation Awareness	.808	.653
Judgment	.799	.638
Flight Leadership – Conduct	.782	.612
Briefed Objectives Fulfilled	.673	.452
Overall Engagement Grade	.802	.644
Eigenvalue	24.99	
Percent of Variance	62.47	